

**LOGIT MODELS FOR ESTIMATING  
URBAN AREA THROUGH TRAVEL**

A Thesis

by

ERIC TALBOT

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

**MASTER OF SCIENCE**

August 2010

Major Subject: Civil Engineering

**LOGIT MODELS FOR ESTIMATING  
URBAN AREA THROUGH TRAVEL**

A Thesis

by

ERIC TALBOT

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee, Mark Burris  
Committee Members, Luca Quadrifoglio  
Michael Sherman  
Head of Department, John Niedzwecki

August 2010

Major Subject: Civil Engineering

## **ABSTRACT**

Logit Models for Estimating Urban Area Through Travel. (August 2010)

Eric Talbot, B.S., Brigham Young University

Chair of Advisory Committee: Dr. Mark Burris

Since through trips can be a significant portion of travel in a study area, estimating them is an important part of travel demand modeling. In the past, through trips have been estimated using external surveys. Recently, external surveys were suspended in Texas, so Texas transportation planners need a way to estimate through trips without using external surveys. Other research in the area has focused on study areas with a population of less than 200,000, but many Texas study areas have a population of more than 200,000. This research developed a set of two logit models to estimate through trips for a wide range of study area sizes, including larger study areas. The first model estimates the portion of all trips at an external station that are through trips. The second model distributes those through trips at one external station to the other external stations. The models produce separate results for commercial and non-commercial vehicles, and these results can be used to develop through trip tables. For predictor variables, the models use results from a very simple gravity model; the average daily traffic (ADT) at each external station as a proportion of the total ADT at all available external stations; the number of turns on the routes between external station pairs; and whether the route is valid, where a valid route is one that passes through the study area and does not pass through any other external stations. Evaluations of the performance of the models showed that the predictions fit the observations reasonably well; at least 68 percent of the absolute prediction errors for each model and for the

models combined were less than 10 percent. These results indicate that the models can be useful for practical applications.

To Breanne

## **ACKNOWLEDGEMENTS**

I would like to thank my committee chair, Dr. Burris for his guidance and support throughout the course of this research, and my committee members, Dr. Quadrifoglio and Dr. Sherman for their valuable comments. I would also like to thank Steve Farnsworth, David Pearson and Ed Hard of the Texas Transportation Institute for their help with obtaining and understanding the data used in this research.

## TABLE OF CONTENTS

	Page
ABSTRACT .....	iii
DEDICATION .....	v
ACKNOWLEDGEMENTS .....	vi
TABLE OF CONTENTS .....	vii
LIST OF FIGURES.....	ix
LIST OF TABLES .....	x
 CHAPTER	
I      INTRODUCTION.....	1
II      LITERATURE REVIEW .....	4
The Modlin and Pigman Models .....	4
Regional Context Models.....	7
Patterns and New Developments.....	11
III     RESEARCH OBJECTIVE AND APPROACH.....	15
Through Trip Estimation Process.....	15
Survey Data .....	20
Proposed System of Models .....	21
IV     THROUGH TRIP SPLIT MODEL DEVELOPMENT AND EVALUATION .....	25
Candidate Predictor Variables.....	25
Preliminary Variable Selection .....	38
Diagnostics .....	43
Final Variable Selection .....	48
Model Refinement.....	50
Model Evaluation .....	52
Results .....	57

CHAPTER		Page
V	THROUGH TRIP DISTRIBUTION MODEL DEVELOPMENT AND EVALUATION.....	62
	Candidate Predictor Variables.....	62
	Preliminary Variable Selection .....	70
	Final Variable Selection .....	73
	Model Refinement.....	75
	Model Evaluation .....	77
	Results .....	81
	Combined Model Evaluation .....	83
VI	CONCLUSIONS .....	87
	REFERENCES .....	91
	VITA .....	94



## LIST OF FIGURES

FIGURE	Page
1    Interaction of the through trip split and distribution models.....	22
2    Local route validity example.....	34
3    Average turns example.....	35
4    Forward selection criteria versus number of variables for the through trip split model .....	42
5    Model Ia diagnostics .....	45
6    Observed through trip probability versus number of responses.....	47
7    Step test for $PINTTH_j$ .....	51
8    Box plots of the split model prediction errors for each quartile of the prediction values .....	56
9    A well-behaved study area .....	58
10   An ill-behaved study area.....	59
11   Through route validity example .....	68
12   Forward selection criteria versus number of variables for the through trip distribution model .....	72
13   Step test for model IIb continuous variables .....	76
14   Box plots of the distribution model prediction errors for each quartile of the prediction values (for observations equal to zero) .....	80
15   Box plots of the distribution model prediction errors for each quartile of the prediction values (for observations not equal to zero) .....	81
16   Box plots of the combined models prediction errors for each quartile of the prediction values .....	86

## LIST OF TABLES

TABLE	Page
1 Variables for Stage One Models from Previous Research .....	12
2 Variables for Combined Models from Previous Research .....	13
3 Variables for Stage Two Models from Previous Research .....	14
4 Survey Data Study Areas .....	21
5 Candidate Variables for the Through Trip Split Model .....	26
6 Interaction Score Definition .....	31
7 Forward Selection Results for the Through Trip Split Model .....	41
8 Model Ia .....	43
9 Model Ia with Observation 637 Removed .....	46
10 Study Area Groups .....	49
11 Model Ia Relative Parameter Estimate Changes .....	49
12 Model Ib .....	50
13 Model Ic .....	52
14 Predicted versus Observed Percent Through Trips for Group 1 .....	53
15 Predicted versus Observed Percent Through Trips for Group 2 .....	53
16 Predicted versus Observed Percent Through Trips for Group 3 .....	54
17 Predicted versus Observed Percent Through Trips for Group 4 .....	54
18 Summary of Through Trip Split Model Prediction Errors .....	55
19 Candidate Predictor Variables for the Through Trip Distribution Model..	63

TABLE		Page
20	Forward Selection Results for the Through Trip Distribution Model.....	71
21	Model IIa .....	73
22	Model IIa Relative Parameter Estimate Changes.....	74
23	Model IIb.....	74
24	Model IIc .....	76
25	Predicted versus Observed Through Trip Distribution for Group 1 .....	78
26	Predicted versus Observed Through Trip Distribution for Group 2 .....	78
27	Predicted versus Observed Through Trip Distribution for Group 3 .....	78
28	Predicted versus Observed Through Trip Distribution for Group 4 .....	79
29	Summary of Through Trip Distribution Model Prediction Errors .....	79
30	Combined Model Predicted versus Observed for Group 1 .....	84
31	Combined Model Predicted versus Observed for Group 2 .....	84
32	Combined Model Predicted versus Observed for Group 3 .....	85
33	Combined Model Predicted versus Observed for Group 4 .....	85
34	Summary of Combined Model Prediction Errors .....	85

## CHAPTER I

### INTRODUCTION

Through trips, or trips that pass through a study area with both their origin and destination outside the study area, are an important part of travel demand modeling. To develop and calibrate the through-trip component of travel demand models, Metropolitan Planning Organizations (MPOs) use data from external surveys, which gather information from travelers entering or leaving the study area. In the past, transportation planners conducted external surveys using the road-side interview technique at locations (called external stations) where traffic enters and exits the study area. During the daylight hours of a certain day, survey personnel would direct all vehicles or a sample of vehicles leaving the urban area to stop on the highway shoulder. The survey personnel would then ask the drivers for information, including the origin, destination, and purpose of the driver's trip. After completing the survey, the survey personnel would allow the drivers to continue on their trips. Transportation planners would then use the information from the drivers to estimate through trip travel patterns.

Although the roadside interview method is an effective way to collect information on through trip travel patterns, it also has potential drawbacks. First, the roadside interview method may create an unsafe situation for drivers because they may not expect to encounter stopped or slowed traffic at the location where the survey is conducted. Second, roadside interviews cause delays for drivers who are surveyed, and may also cause delays for drivers who are not surveyed. Third, some drivers resent being stopped for a survey as an invasion of their privacy. Fourth, the roadside interview method is expensive, with the cost for a complete set of external surveys for a study area usually exceeding \$100,000.

For some or all of these reasons, in early 2008 the Texas Department of Transportation suspended external surveys throughout the state. Texas MPOs now need a way to estimate through trips without recent external survey data. Previous research has developed models for estimating through trip patterns, but most of these models

focus on study areas with less than 50,000 people, while the study areas for MPOs always have more than 50,000 people. In addition, most of the previous research used linear regression, which may not be the best approach because the variable of interest is a proportion, rather than a continuous number.

To improve on these previous models, the research described in this document developed a system of two models. Each model was developed using logistic regression, which is appropriate for data where the responses are proportions. The first model is a binary logit model which estimates the proportion of trips at an external station that are through trips. The second model is a multinomial logit model which distributes the through trips between all the external stations where a through trip could have entered the study area. These models were developed for urban areas with 50,000 to 6 million people, so they are applicable to all but one of the MPO study areas in Texas. Model evaluation shows that these models perform well and will be useful for estimating through travel in Texas and other areas.

This document describes the research process that led to developing the new model system, presents the results, and discusses the results' meaning and significance. The research process started with a literature review, which found much useful information from previous research, but also identified the limitations in previous research discussed earlier which would limit its application to study areas across Texas. Based on these findings, the objective of the research became to develop a method to estimate through trips that would be applicable to larger urban areas. Then an appropriate research approach was developed, using as guides the research objective, the findings from previous research, the kind of data available, and preliminary data analysis. Then a system of two logit models was developed, and each model was evaluated. The research methods used to carry out the research approach included common statistical model fitting and evaluation procedures. The literature review, research objective and approach, research methods and results, and research conclusions are detailed in the following chapters, outlined below:

- Chapter II: Literature Review

- Chapter III: Research Objective and Approach
- Chapter IV: Through Trip Split Model Development and Evaluation
- Chapter V: Through Trip Distribution Model Development and Evaluation
- Chapter VI: Conclusions

## **CHAPTER II**

### **LITERATURE REVIEW**

The economic and social conditions of an urban area are greatly affected by the quality of the area's transportation system. Transportation planners seek to improve long-term economic and social conditions by effectively coordinating and programming future transportation system investments. One of the most important methods transportation planners use is estimating current travel patterns in the urban area, then predicting future travel patterns using a travel demand model. These predictions can then be used to determine future transportation needs.

An accurate understanding of the travel patterns in an urban area requires an accurate estimate of through-trip travel patterns. Through trips often form a significant portion of all travel within a study area. For example, data from a 2005 study suggests that through trips account for 17 percent of all vehicle-miles traveled within the Austin, Texas study area (S. Farnsworth, unpublished project proposal, 2009). The portion is usually even higher for smaller study areas.

Because external surveys are not currently performed in Texas, transportation planners need another method to estimate through-trips. Developing such a method is the goal of this research. The research process began with a literature review covering research on methods to estimate through trips without an external survey. This section presents a historical summary of the literature, then discusses general patterns found in the research.

#### **THE MODLIN AND PIGMAN MODELS**

The earliest through trip estimation method reviewed was created by Modlin, who developed a set of multiple linear regression equations to estimate through trips for small study areas (study areas with less than 50,000 people). He used the roadway functional classification, average daily traffic, percent trucks, and percent pickups and panel vans at each external station, and the population of the study area as explanatory variables. Modlin's model has two stages. The first stage estimates the percent of all trips at each external station that are through trips. The second stage distributes the through

trips between external stations (Modlin 1974). A few years later, Pigman published a set of similar equations for small study areas (Pigman and Deen 1979), and Modlin followed up with a new set of regression equations in 1982. This second set of equations was similar to his first, but also included route continuity, which is a binary variable signifying whether or not two external stations are on the same highway route (Modlin 1982).

As an example of these early models, Modlin's 1982 stage one model is

$$Y = 9.29 - 0.00031UP + 0.0026ADT + 1.48TRK$$

where

$Y$  = percentage of through-trip ends of the ADT at the external station;

$UP$  = urban area population;

$ADT$  = average daily traffic at the external station; and

$TRK$  = percentage of trucks excluding panels and pickups at the external station.

The stage two equation for interstates is

$$Y = -2.70 + 0.21PTTDES + 67.86RTECON;$$

for principal arterials it is

$$Y = -7.40 + 0.55PTTDES + 24.68RTECON + 45.62ADT / CD;$$

for minor arterials it is

$$Y = -0.63 + 86.68ADT / CD + 30.04RTECON;$$

for major collectors it is

$$Y = -1.08 + 0.00079DESADT + 0.47PTKDES + 31.78ADT / CD;$$

and for minor collectors and local roads it is

$$Y = -0.40 + 109.42ADT / CD$$

where

$Y$  = percentage distribution of through-trip ends from an origin station to a destination station;

$PTTDES$  = percentage of estimated through-trip ends at the destination station;

$RTECON$  = route continuity (1 = yes, 0 = no);



$ADT / CD =$  ADT at destination station divided by the sum of ADT at all stations;

$DESADT =$  ADT at destination station; and

$PTKDES =$  percentage trucks excluding panels and pickups at the destination station.

Chatterjee compared the two Modlin models and the Pigman model using external origin destination-data from 14 small study areas in North Carolina. He found that the 1974 Modlin model produced the best results, but also required the greatest data collection efforts. All three of the models performed reasonably well for external stations with high traffic volumes, but performed erratically for external stations with low traffic volumes (Chatterjee and Raja 1989). Reeder tested the 1982 Modlin model and the Pigman model to determine if they could be applied to medium-sized study areas (study areas with 50,000 to 200,000 people), using results from eight external surveys from four study areas in Texas. Both models had large mean square errors and frequently estimated negative numbers of trips when applied to the larger urban areas (Reeder 1993).

In 1998, NCHRP Report 365, "Travel Estimation Techniques for Urban Planning," reprinted some of Modlin's 1974 and 1982 equations, but also stated that,

"[e]xternal travel estimation has been the least documented component of the travel demand models. ... Research into external travel revealed that very little has been done in the advancement of external travel estimation. ... It is recommended that local external travel data be collected to the extent possible and that further research is needed into the collection and estimation of external travel." (Martin and McGuckin 1998)

In response, some more recent researchers have worked in the years since the publication of NCHRP Report 365 on developing newer and better models for estimating through trips. Because previous models considered the study area in isolation

from the rest of the world, much of the work has focused on incorporating the geographic and economic context of the study area into the model. The next section reviews this work.

## REGIONAL CONTEXT MODELS

In a 1999 paper, Anderson reported on the results of an evaluation of the effectiveness of a simple gravity model, Huff's model, and Zipf's model for estimating through trips. Huff originally developed his model to estimate the probability that a customer living at a certain location would patronize a certain shopping center (Huff 1963). Anderson adapted the model to estimate the probability that a trip starting at urban area  $i$  would end at urban area  $j$  using

$$p_{ij} = \frac{A_j}{\sum_{q \in U} A_q}$$

and

$$A_j = \frac{P_j}{D_{ij}^\lambda}$$

where

$p_{ij}$  = the probability that a trip starting at location  $i$  would end at urban area  $j$ ;

$A_j$  = the attractiveness of urban area  $j$ ;

$U$  = the set of urban areas forming the choice set;

$P_j$  = the population of urban area  $j$ ;

$D_{ij}$  = the distance between location  $i$  and urban area  $j$ ; and

$\lambda$  = a model parameter.

For the parameter  $\lambda$  Anderson used 1.

The Zipf model postulates that the volume of travel between two cities is linearly proportional to the product of the two cities' populations divided by the distance between

the two cities (Zipf 1946). Anderson adapted the model to estimate the volume of travel between cities using

$$I_{ij} = k \frac{P_i P_j}{D_{ij}^\lambda}$$

where

$I_{ij}$  = the interaction between urban areas  $i$  and  $j$ ;

$k$  = a model parameter;

$P_i, P_j$  = the populations of urban areas  $i$  and  $j$ ;

$D_{ij}$  = the distance between urban areas  $i$  and  $j$ ; and

$\lambda$  = a model parameter.

For the parameters  $k$  and  $\lambda$ , Anderson used 2 and 0.6, based on fitting the model to a calibration data set.

Anderson tested the three models using a city in Iowa with four external stations and a population of 8500. Anderson compared the results from these three models to results from the Modlin regression equations, and to observed through trip patterns, and found that Huff's model and the Modlin regression equations both estimated the observed values reasonably well (Anderson 1999). Anderson later applied Huff's model to three small cities in Alabama, and found that the spatial-economic model estimated observed through trips well (Anderson 2005).

Horowitz developed a model which assigns a "catchment" area from the region outside of the study area to each external station, and calculates a weight factor for trips between two external stations by calculating the probability that a line connecting two points within their catchment areas passes through the study area, or crosses a barrier to travel between catchment areas. These weight factors are then used to estimate through trips using the procedure outlined in the first Quick Response Freight Manual (Cambridge Systematics Inc. 1996). For one of two test urban areas, the new model explained 96 percent of the variation in the through trip patterns, compared to 88 percent for a model with all weight factors set to 1. For the second test urban area, the new

model explained 99.9 percent of the variation, compared to 99.4 percent for the “all 1” model (Horowitz and Patel 1999).

Han combined the work of Modlin, Anderson and Horowitz. Like the Modlin model, Han's model is a set of regression equations, and uses information about the traffic and roadway at the external station, and about the study area, as explanatory variables. However, the Han model also includes Zipf probabilities and Horowitz weights as explanatory variables. Han's work differed from the previous work in that the model was developed small and medium sized study areas with up to 200,000 people (Han 2007; Han and Stone 2008).

Han's stage one model is

$$Y = (3.353 - 0.850Other + 1.671Small + 2.682MR + 0.000104ADT - 0.000029Pop + 0.046TRK + 0.0012Area + 0.000026Emp)^2$$

where

$Y$  = percentage of through trip ends of ADT at the external station;

$Other$  = 1 if the external station is a collector or local road, and 0 otherwise;

$Small$  = 1 if the urban area has a population of less than 50,000 people, and 0 otherwise;

$MR$  = 1 if the external station is on a marginal route, and 0 otherwise;

$ADT$  = average daily traffic;

$Pop$  = population of the study area;

$TRK$  = percentage of trucks at the external station;

$Area$  = area size of the study area in square miles; and

$Emp$  = employment in the study area.

Han's stage two model for study areas with a population of less than 50,000 is

$$Y_{ij} = (1.42 + 1.29RTECON + 0.73D\_LANE - 0.03D\_PTT + 2.00Probl + 1.64D\_Zipf)^2$$

and his model for study areas with a population between 50,000 and 200,000 is

$$Y_{ij} = (0.20 + 5.03RTECON + 0.19D\_ADT\_CD + 1.13Prob3 - 0.04O\_PTT)^2,$$

where

$Y_{ij}$  = percentage distribution of through trip ends from origin station  $i$  to destination station  $j$ ;

$RTECON$  = 1 if external stations  $i$  and  $j$  are on the same continuous highway route;

$D\_LANE$  = number of highway lanes at the destination station;

$D\_PTT$  = percentage through trip ends at the destination station, as estimated by the stage one model;

$O\_PTT$  = percentage through trip ends at origin station, as estimated by the stage one model;

$Prob1$  = Horowitz's weight for external stations  $i$  and  $j$  when the width of the catchment area is one-quarter of the simulated study area radius;

$Prob3$  = Horowitz's weight for external stations  $i$  and  $j$  when the width of the catchment area is three-quarters of the simulated study area radius;

$D\_Zipf$  = Zipf's probability factor for the destination station; and

$D\_ADT\_CD$  = ratio of ADT at destination station to the sum of ADT at all stations.

Han tested the models using a validation data set consisting of external survey results from several study areas. For the stage one model, the root mean square error (RMSE) for each study area ranged from 6.4 to 11.9 percent. For the stage two model, the RMSE was between 4.8 and 28.1 percent.

Anderson also developed Modlin-like regression equations, and included a variable to signify the presence of a near-by major city (Anderson et al. 2006).

## **PATTERNS AND NEW DEVELOPMENTS**

The literature review revealed some general patterns in previous researchers' approach to estimating through trip patterns that helped create a starting point for this research. First, most of the models were some type of statistical regression model. Five of the models used linear regression. Only the Horowitz model and the three spatial economic-models tested by Anderson are non-statistical models. Second, of the five regression models, four were actually a system of two models, where one model (the stage one model) predicted the proportion through trips at an external station, and the second model (the stage two model) distributed the through trips between external stations. The third pattern has already been discussed, which is that the earlier models did not include variables to account for the regional context of the study area, and the later models sought to improve predictive power by including such variables. These patterns served as a guide when forming the research approach, as discussed later in this document.

One other research project also influenced the research approach. This was the work of Martchouk and Fricker, who recently proposed modeling through trips using logistic regression rather than linear regression, as all previous regression-based models had done (Martchouk and Fricker 2009). They developed a logistic model which uses a set of variables similar to those of the Modlin and Pigman models, but which is a one-stage model rather than a two-stage model.

Table 1, Table 2 and Table 3 summarize the results of previous research. Table 1 lists the variables that were considered for at least one stage one model. A check mark indicates that the variable was included in the final model. Table 2 and Table 3 present the same information for combined models and stage two models. These results served as a starting point for choosing the candidate predictor variables for this research.

**Table 1. Variables for Stage One Models from Previous Research**

	Modlin 1974	Pigman 1979	Modlin 1982	Han 2008
Characteristics of the survey station				
Functional Classification	✓			✓
Number of Lanes				
Average Daily Traffic	✓	✓	✓	✓
Percent heavy trucks	✓	✓	✓	✓
Percent pickups and vans	✓			
Zipf's probability factor				
Huff's probability factor				
Marginal highway route				✓
Characteristics of the study area				
Population	✓	✓	✓	✓
Employment				✓
Income				
Surface area				✓
Variables that appear in this table were considered in previous research. A check mark indicates that the variable was included in the researchers' final model.				

**Table 2. Variables for Combined Models from Previous Research**

	Anderson 2006	Martchouk 2009
Characteristics of the survey station		
Average daily traffic		
Percent trucks		
Number of lanes		
Near-by major city	✓	
Characteristics of the choice (entry) station		
Average daily traffic	✓	✓
Percent trucks		
ADT as a portion of total ADT		✓
Number of lanes		
Functional classification		
Near-by major city		
Characteristic of external station pairs		
Route continuity	✓	✓
Characteristics of study area		
Population		
Employment		
Miscellaneous		
Internal-external factor	✓	✓
Variables that appear in this table were considered in previous research. A check mark indicates that the variable was included in the researchers' final model.		



**Table 3. Variables for Stage Two Models from Previous Research**

	Modlin 1974	Pigman 1979	Modlin 1982	Han 2008
Characteristics of the survey station				
Functional classification	✓	✓	✓	
Number of lanes				
Average daily traffic				
ADT as a portion of total ADT				
Zipf's probability factor				
Huff's probability factor				
Stage one results	✓			
Marginal highway route				
Characteristics of the choice (entry) station				
Functional classification	✓			
Number of lanes				✓
Average daily traffic	✓	✓	✓	
Percent heavy trucks	✓	✓	✓	
Percent pickups and vans				
ADT as a portion of total ADT		✓	✓	✓
Zipf's probability factor				✓
Huff's probability factor				
Stage one results	✓	✓	✓	✓
Marginal highway route				
Characteristics of external station pairs				
Route continuity			✓	✓
Angle between stations				
Horowitz's weight				✓
Characteristics of the study area				
Population	✓			✓
Employment				
Income				
Surface Area				
Variables that appear in this table were considered in previous research. A check mark indicates that the variable was included in the researchers' final model.				

### **CHAPTER III**

#### **RESEARCH OBJECTIVE AND APPROACH**

The previous chapter summarized models for through trip estimation that were developed in previous research. A major limitation of these through trip models is that most were developed for study areas with less than 50,000 people, while every Texas study area with a travel demand model has a population of more than 50,000. Between the Modlin, Pigman, Anderson, Horowitz and Martchouk models, all but one of the study areas used for model development had less than about 50,000 people. (The one exception occurred in the Horowitz model, which was tested on a single study area with roughly 150,000 people.) Only the Han model used multiple study areas with more than 50,000 people. Even his model did not include any study areas with more than 200,000 people, while Texas has 13 study areas with more than 200,000 people, each of which maintains a travel demand model. Therefore, the objective of this research is to develop a new method for estimating through trips that is applicable to Texas study areas.

This chapter describes the research approach that was used to achieve the research objective. An understanding of the research approach requires an understanding of both the end product of through trip estimation, and the process that leads to the end product. With this understanding, the challenges caused by eliminating external surveys are obvious. With this motivation, this chapter describes the desired end product of through trip estimation, which is a set of two through trip tables, and explains each of the three steps of the through trip estimation process. An understanding of the research approach also requires a knowledge of the type of data that was available, as this largely controls the types of analysis that are possible. Therefore, this chapter also describes the data that was used for the analysis. Following these explanations, this chapter concludes by presenting and rationalizing the research approach.

#### **THROUGH TRIP ESTIMATION PROCESS**

This section describes through trip tables and the through trip estimation process and how it is affected by eliminating external surveys. However, before proceeding, this

section first presents definitions of three other types of study area trips which will be important to understanding the through trip estimation process.

### **Other Types of Trips in an Urban Area**

Through trips are one of four types of trips in a study area. As explained previously, through trips have both origin and destination outside the study area, but they pass through the study area. Another term for through trip is external-external (E-E) trip. Internal-internal (I-I) trips have both origin and destination inside the study area. External-internal (E-I) trips have an origin outside the study area and a destination inside the study area. Internal-external (I-E) trips have an origin inside the study area and a destination outside the study area. This proposal uses the terms through, I-I, E-I, and I-E for each of the four types of trips. E-I and I-E trips will often be treated as one group, where they are referred to as E-I/I-E trips.

### **Through Trip Tables**

The end product of through trip estimation is a set of two estimated through trip tables, one for commercial vehicles, and one for non-commercial vehicles. A through trip table is a square matrix where each element  $t_{ij}$  is the average number of through vehicle trips entering the study area at  $i$  and leaving the study area at  $j$ . The indices  $i$  and  $j$  refer to “external stations,” or places where traffic crosses the study area boundary. If a study area has  $S$  external stations, then the dimensions of the through trip table are  $S \times S$ . The elements on row  $i$  represent trips entering the study area at external station  $i$ , and the elements in column  $j$  represent trips leaving the urban area at external station  $j$ . The elements of the through trip table usually represent the number of vehicle trips on an average weekday 24 hour period. The elements on the main diagonal, where  $i = j$ , are zero, as through trips would not enter and leave the study area at the same external station. In addition, the through trip matrix is usually assumed to be symmetrical about the main diagonal, which means that the number of trips from  $i$  to  $j$  is equal to the number of trips from  $j$  to  $i$  over a 24 hour period.

A through trip table is the product of a three-step through trip estimation process, described in the next section.

### **Through Trip Estimation Process: Steps**

The through trip estimation process has three steps. The three steps are:

- for each external station, count the number of vehicle trips and estimate the proportion of all trips that are commercial vehicle trips;
- for each external station, and for each vehicle type (commercial and non-commercial), estimate the proportion of vehicles trips that are E-I/I-E trips, and the proportion that are through trips coming from or going to each of the other external stations; and
- develop the through trip tables based on information from Steps 1 and 2.

This section describes each step for the case when an external survey is available, and examines how each step is (or is not) affected by the absence of external surveys. This section also proposes, in a very general way, solutions for problems that arise when an external survey is not available.

#### *Through Trip Estimation Process Step 1*

Step 1 involves counting the total number of vehicles entering and exiting the study area at each external station. The counts usually occur simultaneously at all the external stations during one 24-hour weekday period, using automated counting machines, such as pneumatic tubes counters, inductive loop detectors, or video detection. The counts serve as an estimate of the average weekday traffic volume at each external station.

Step 1 also involves estimating the proportion of vehicles that are commercial vehicles. A commercial vehicle is any vehicle being used for a commercial purpose, and could be a passenger car, pickup truck, or van. However, commercial vehicles are almost always larger vehicles, such as moving vans and tractor-trailer combinations. For this reason, the proportion of larger vehicles in the count is usually used as an estimate of the proportion of commercial vehicles in the sample. Most automated counting machines are

capable of classifying vehicles according to vehicle size, and are used to estimate the proportion of larger vehicles.

If external surveys are part of the estimation process, then they usually occur at the same time and place as the counts. However, the counts are not expensive, can still be done in Texas, and are easily implemented whether or not they are paired with an external survey. Thus eliminating external surveys has no effect on Step 1, and Step 1 will not be a focus of this research.

#### *Through Trip Estimation Process Step 2*

The four types of trips in a study area are through trips, I-I trips, and E-I trips and I-E trips. Through trips cross the study area boundary twice; they enter the study area at one external station and exit the study area at another external station. E-I/I-E trips cross the study area boundary once, either entering or exiting the urban area at one external station. I-I trips never pass through an external station, so the counts at each external station from Step 1 serve as an estimate of the sum of through trips and E-I/I-E trips at each external station. However, the counts do not distinguish between each type of trip, nor do they identify the entry external station for each through trip leaving the study area. Thus the role of Step 2 is to estimate the proportion of all trips at each external station, for each vehicle type, that are through trips, (the through-trip split), and the proportion of these through trips that are coming from or going to each of the other external stations (the through-trip distribution).

The estimation problem is simplified with the assumption that the true through trip table is symmetrical about the main diagonal. Then an estimate for the proportions among outbound vehicles also serves as an estimate for the proportions among inbound vehicles. In Texas, the external survey-based estimation method has used this assumption, with external surveys almost always administered to outbound drivers only.

The through trip split and distribution can be estimated using an external survey. To conduct an external survey, surveyors position themselves on the roadside at each external station, and direct vehicles to pull to the side of the road and stop. Surveyors then interview the drivers and collect information such as the origin and destination of

the trip, the entry external station for through trips, and the vehicle type. After the interview is completed, the surveyors allow the drivers to continue on their trip. This continues throughout the day, with the surveyors stopping vehicles and allowing vehicles to pass through as needed to obtain an adequate sample size and ensure sampling continues at a regular rate throughout the day. Once the external survey is complete, the through trip split and distribution are estimated using the sample proportions calculated from the survey responses.

Obviously, eliminating the external survey will completely change the procedures for Step 2. Without a survey, there is a need for another way to estimate the through trip split and distribution, so Step 2 will be a focus of this research.

### *Through Trip Estimation Process Step 3*

Step 3 uses the information from Steps 1 and 2 to develop an estimated through trip table. The through trip table must somehow agree with the counts from Step 1 and the proportions from Step 2 while satisfying the assumption that the table is symmetric about the main diagonal. A variety of methods exist for solving this problem, from very simple to very computationally and theoretically complex (for an example of a moderately complex method, see Spiess [1987]). A method often used in Texas estimates the through trip tables using

$$t_{com,ij} = (1/4) \cdot (p_{split,com,j} \cdot p_{dstr,com,ij} \cdot ADTLV_j + p_{split,com,i} \cdot p_{dstr,com,ji} \cdot ADTLV_i)$$

and

$$t_{non,ij} = (1/4) \cdot (p_{split,non,j} \cdot p_{dstr,non,ij} \cdot ADTSV_j + p_{split,non,i} \cdot p_{dstr,non,ji} \cdot ADTSV_i)$$

where

$t_{com,ij}$  = the number of through trips entering the study area at external station  $i$  and exiting the study area at external station  $j$ , for commercial vehicles;

$t_{non,ij}$  = the number of through trips entering the study area at external station  $i$  and exiting the study area at external station  $j$ , for non-commercial vehicles;

$p_{split,com,i}$  = the portion of commercial vehicle trips exiting the study area at external station  $i$  that are through trips (the through trip split for commercial vehicles);

$p_{dstr,com,ij}$  = the portion of commercial vehicle through trips exiting the study area at external station  $j$  that entered the study area at external station  $i$  (the through trip distribution for commercial vehicles);

$p_{split,non,i}$  = the portion of non-commercial vehicle trips exiting the study area at external station  $i$  that are through trips (the through trip split for non-commercial vehicles);

$p_{dstr,com,ij}$  = the portion of non-commercial vehicle through trips exiting the study area at external station  $j$  that entered the study area at external station  $i$  (the through trip distribution for non-commercial vehicles);

$ADTLV_i$  = the two-way ADT for large vehicles at external station  $i$  (from step 1); and

$ADTSV_i$  = the two-way ADT for small vehicles at external station  $i$  (from step 1).

When an external survey is available, this step uses the proportions estimated using the external survey to estimate the through trip tables. If the proportions are estimated using a different method, as has been suggested, then this step may need to be modified to use the best data available.

## **SURVEY DATA**

The research approach is largely controlled by the data that is available. The data for this research comes from external surveys performed in 13 study areas in Texas between 2001 and 2006. Table 4 includes the name of each study area, the year that the survey was performed, the number of external stations that were surveyed, and the total number of survey responses. The survey responses are the response variables for the

model developed by this research. Details on the data and data sources for the predictor variables are given in the next two chapters.

**Table 4. Survey Data Study Areas**

Study Area	Year	Survey External Stations	Total Responses
Abilene	2005	11	3329
Amarillo	2005	12	4234
Austin	2005	22	8298
Dallas – Fort Worth	2005	32	12642
Longview	2004	30	8426
Lubbock	2005	17	3988
Midland – Odessa	2002	13	4023
San Angelo	2004	11	4031
San Antonio	2005	22	9892
Sherman – Denison	2005	10	3975
Tyler	2004	18	5124
Waco	2006	15	4557
Wichita Falls	2005	11	3093

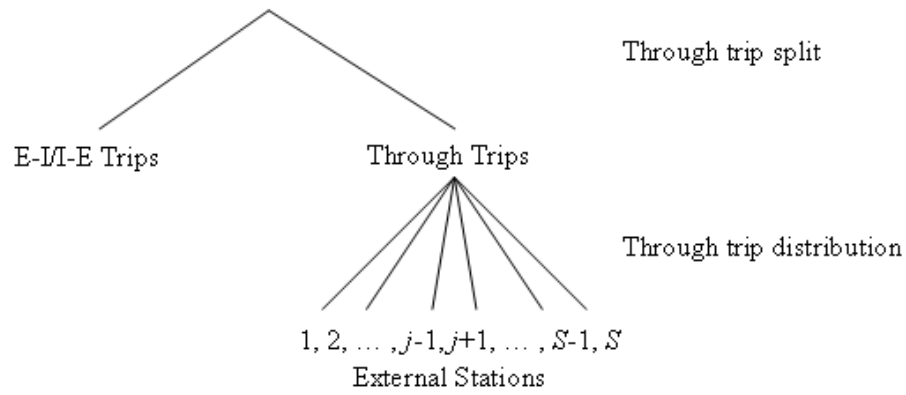
## PROPOSED SYSTEM OF MODELS

To achieve the research objective, this research developed two models. The first model (the through trip split model) estimates the proportion of trips exiting the study area at an external station that are through trips. The remaining proportion is the proportion of trips that are E-I/I-E trips. The second model (the through trip distribution model) estimates the proportion of through trips exiting the study area at an external station that entered the study area at each of the other external stations. Multiplying the result for each entry external station from the through trip distribution model by the result from the through trip split model estimates the proportion of all trips that are through trips that entered at each of the other external stations.

Figure 1 illustrates how the two models would work together to estimate the through trip split and distribution for an external station  $j$ . The upper part of the illustration shows the through trip split model estimating the proportion of trips at



external station  $j$  that is through trips. The lower part shows the through trip distribution model estimating the proportion of through trips that entered at each of the possible entry external stations, where  $S$  is the number of external stations in the study area. No estimation is made for external station  $j$ , since through trips would not enter and exit the study area at the same external station.



**Figure 1. Interaction of the through trip split and distribution models.**

The through trip split and distribution models are both logit models. Logit models are appropriate when the response is one of a finite number of outcomes (Hosmer and Lemeshow 2000). The through trip split model is a binary logit model, where the response has two possible outcomes (E-I/I-E or through). The through trip distribution model is a multinomial logit model, where the response has three or more possible outcomes (each of the possible entry external stations). For the through trip split model, the through response is  $Y = 1$ , and the E-I/I-E response is  $Y = 0$ . The through trip split model has the form

$$P_j(Y = 1) = \frac{\exp[g(\mathbf{x}_j)]}{1 + \exp[g(\mathbf{x}_j)]}$$

where  $g(\mathbf{x}_j) = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \dots + \beta_p x_{jp}$ ,  $p+1$  is the number of parameters in the model,  $j$  indexes the external station for which the estimation is being made, and  $P_j(Y=1)$  is the probability of a through trip at external station  $j$ .

For the through trip distribution model, one of the responses is the baseline response, coded as  $Y=1$ . The other responses are  $Y=2,3,\dots,j-1,j+1,\dots,S-1,S$ , where  $S$  is the number of external stations in the study area, and  $S-1$  is the number of possible responses (since a through trip cannot enter and exit the study area at the same external station). The through trip distribution model has the form

$$P_j(Y=s) = \frac{\exp[g(\mathbf{x}_{sj})]}{\sum_{q=1,2,\dots,j-1,j+1,\dots,S-1,S} \exp[g(\mathbf{x}_{jq})]}$$

where  $g(\mathbf{x}_{sj}) = \beta_1 x_{sj1} + \beta_2 x_{sj2} + \dots + \beta_p x_{sjp}$ ,  $p$  is the number of parameters in the model,  $s$  indexes the response (an external station where a through trip can enter the study area),  $j$  indexes the external station for which the estimation is being made (the external station where through trips exit the study area, and  $P_j(Y=s)$  is the probability that a through trip that exits the study area at external station  $j$  entered the study area at external station  $s$ . The specification of  $g(\mathbf{x}_{sj})$  has no alternative specific constant because there is only one type of alternative: external stations.

For the binary logit model, the parameter estimates result from maximizing

$$l(\beta) = \prod_{n=1,\dots,N} [P_j(Y=0)]^{y_{0n}} [P_j(Y=1)]^{y_{1n}}$$

where

$n$  = an index for each response;

$N$  = the total number of responses;

$j$  = an index corresponding to the external station where response  $n$  was observed;

$y_{0n}$  = 1 if the response  $n$  is 1, and is 0 otherwise; and

$y_{1n}$  = 1 if the response  $n$  is 0, and is 0 otherwise.

For the multinomial logit model, the parameter estimates result from maximizing

$$l(\beta) = \prod_{n=1, \dots, N} [P_j(Y=1)]^{y_{1n}} [P_j(Y=2)]^{y_{2n}} \dots [P_j(Y=j-1)]^{y_{(j-1)n}} [P_j(Y=j+1)]^{y_{(j+1)n}} \\ \dots [P_j(Y=S-1)]^{y_{(S-1)n}} [P_j(Y=S)]^{y_{Sn}}$$

where

$n =$  an index for each response;

$N =$  the total number of responses;

$j =$  an index corresponding to the external station where response  $n$  was observed; and

$y_{sn} = 1$  if the response  $n$  is  $s$ , and is 0 otherwise.

Most previous research has used linear models instead of logit models to predict through trip proportions. This research uses logit models because they have statistical and practical advantages over linear models. Previous research has fit linear models to the through trip proportions estimated from external surveys. Using the through trip proportions, rather than the number of responses for each possible outcome, results in a loss of all information about sample size. In addition, the linear model can result in proportion predictions that are more than one or less than zero, and can result in estimates that do not sum to one. The logit models retain information about sample size, and the estimated proportions always sum to one as they should, with no estimates greater than one or less than zero.

Another possible model form is the nested logit model, where a single model could replace the through trip split and distribution models. However, preliminary analysis showed erratic and poor results for the nested logit model, probably resulting from the fact that attributes of two very different kinds describe the E-I/I-E outcome and the through outcomes. In addition, even if a good nested model exists for this problem, the model would probably be hard to understand and interpret, again because of the different kinds of attributes for the outcomes.

## **CHAPTER IV**

### **THROUGH TRIP SPLIT MODEL DEVELOPMENT AND EVALUATION**

This research created a system of two models for estimating through trips in study areas. This chapter focuses on the development of the through trip split model which predicts the portion of trips at an external station that are through trips. The model development started with choosing candidate predictor variables. Then a subset of the candidate predictor variables was chosen to form a preliminary model, and the fit of the preliminary model was evaluated using model diagnostics. Then a final variable selection was made, this time from the variables in the preliminary model. The possibility of refining the model composed of the variables from the final selection using transformation and interactions was then investigated. The final model was then evaluated to determine its goodness of fit and practical applicability. This chapter describes each of these steps in detail.

#### **CANDIDATE PREDICTOR VARIABLES**

The model development process started by selecting candidate predictor variables, which are variables that have good potential for predicting through trips and merit further analysis. After defining each of the candidate predictor variables, this section discusses some of the more complicated variables and variables that are new to predicting through trips. This section then explains why some of the variables that were used in previous research were not considered in this research. Finally, this section describes the data source for each of the predictor variables.

#### **Variable Definitions**

Each of the candidate predictor variables for the through trip split model is defined in Table 5. Several of the variables depend on the interaction score, which is defined in Table 6. Following the pattern of previous chapters, the subscript  $j$  refers to the external station for which the through trip estimation is made, also called the survey station.

**Table 5. Candidate Variables for the Through Trip Split Model**  
**Traffic Characteristics**

$ADTALL_j$  = The average daily traffic (ADT) for external station  $j$  for all vehicle types, where ADT is the average non-holiday weekday 24-hour two-way count of vehicles passing through the external station

$ADTLV_j$  = The ADT for external station  $j$  for large vehicles, where large vehicles are vehicles belonging to classes 4 through 13 of the Federal Highway Administration (FHWA) vehicle classification system

$$PROPLV_j = \frac{ADTLV_j}{ADTALL_j}$$

The portion of the total ADT that is large vehicles

$$PADTSV_j = \frac{ADTSV_j}{\sum_{q \in E} ADTSV_q}$$

The small vehicle ADT as a portion of the total small vehicle ADT across all external stations, where  $E$  is the set of all external stations in the study area,  $ADTSV_j$  and  $ADTSV_q$  are the ADTs for small vehicles for external stations  $j$  and  $q$  respectively, and small vehicles are vehicles belonging to classes 1 through 3 of the FHWA vehicle classification system  $ADTSV_j$  was not considered as a candidate variable because it is a linear combination of  $ADTALL_j$  and  $ADTLV_j$ .

Table 5 continued

---


$$PADTLV_j = \frac{ADTLV_j}{\sum_{q \in E} ADTLV_q}$$

The large vehicle ADT as a portion of the total large vehicle ADT across all external stations, where  $PADTLV_j$  and  $PADTLV_q$  are the ADTs for large vehicles for external stations  $j$  and  $q$  respectively

$$PADTAL_j = \frac{ADTALL_j}{\sum_{q \in E} ADTALL_q}$$

The ADT for all vehicles at station  $j$  as a portion of the total ADT for all vehicles across all external stations, where  $ADTALL_j$  and  $ADTALL_q$  are the ADTs for all vehicles for external  $j$  and  $q$  respectively

---

---

**Table 5 continued**  
**Roadway Characteristics**

---

$LANES_j$  = Total number of lanes in both directions at external station  $j$ . For example, the value of  $LANES_j$  for an external station with two lanes in each direction would be 4. The lane count only includes main through lanes. Any turning lanes, median left turn lanes, climbing lanes, frontage road lanes or passing lanes are not counted.

$DIVIDED_j$  = A binary variable which is 1 when, in the area of external station  $j$ :  
(1) the two directions of traffic are separated by either a non-traversable barrier, such as a wall or railing, or by a non-paved area which is not intended for traffic, such as a grassy median; and (2) opportunities for left turns across the barrier or non-paved area at an intersection are less frequent than is typical for an urban arterial. The variable is 0 otherwise.

$LIMITED_j$  = A binary variable which is 1 when the roadway in the area of external station  $j$  is a limited-access facility, which means that access to the roadway is only provided by ramps. For areas where the roadway transitions from limited access to non-limited access the variable is 1. The variable is 0 otherwise.

---

---

**Table 5 continued**  
**Interaction Score Variables**

---

$$INTTHR_j = \sum_{\{q \in E, q \neq j\}} INT_{qj}$$

The total interaction score for through trips, where  $INT_{qj}$  is the through interaction score for entry external station  $q$  and survey external station  $j$ , as defined in Table 6

$$INTTTL_j = INTTHR_j + INTLCL_j$$

The total interaction score for all trips, where  $INTLCL_j$  is the interaction score for E-I/I-E (local) trips for survey station  $j$ , as defined in Table 6

$$PINTTH_j = \frac{INTTHR_j}{INTTTL_j}$$

The interaction score for through trips as a portion of the interaction score for all trips. If  $INTTTL_j = 0$  then  $PINTTH_j = 0$

$INTTTL_j =$  A binary variable which indicates if the total interaction score for all trips ( $INTTTL_j$ ) is greater than zero. It is 1 if  $INTTTL_j$  is greater than 0, and is 0 otherwise

$PINTTH_j =$  A binary variable which indicates if the interaction score for through trips as a portion of the interaction score for all trips ( $PINTTH_j$ ) is greater than zero. It is 1 if  $PINTTH_j$  is greater than 0, and is 0 otherwise

---



**Table 5 continued**  
**Route Validity**

$RTELCL_j$  = A binary variable which is 1 if the route from the centroid of at least one U.S. Census urban area or urban cluster whose centroid is inside the study area to external station  $j$  is valid, and is 0 otherwise. A route is valid if (1) it passes through the study area and (2) it crosses the study area boundary only at external station  $j$ . The route is chosen to minimize the travel time under non-congested conditions.

---

**Characteristics of the Study Area**

$POP$  = Population of the study area  
 $EMP$  = Employment in the study area  
 $INC$  = Average income of residents of the study area  
 $AREA$  = Surface area of the study area in square miles  
 $ADTALL$  = The ADT for all vehicles summed across all external stations  
 $ADTLV$  = The ADT for large vehicles summed across all external stations

---

**Is Commercial Vehicle**

$ISCV$  = A binary variable which is 1 if the vehicle is a commercial vehicle, and is 0 otherwise. Here a commercial vehicle is any vehicle used for a commercial purpose, regardless of size or type of vehicle.

---

**Average Turns**

$$AVGTRN_j = \sum_{\{q \in E, q \neq j\}} \left( \frac{DSTR_{qj} \cdot ROUTE_{qj} \cdot TURNS_{qj}}{\sum_{\{r \in E, r \neq j\}} DSTR_{rj} \cdot ROUTE_{rj}} \right)$$

where  $DSTR_{qj}$  and  $DSTR_{rj}$  are the results from the through trip distribution model, and  $ROUTE_{qj}$ ,  $ROUTE_{rj}$ , and  $TURNS_{qj}$  are defined in Table 19;  $j$  is the survey external station; and  $q$  and  $r$  are entry external stations

---

**Table 6. Interaction Score Definition**

$$INT_{ij} = \sum_{\{v \in U, v \neq w\}} \left( \sum_{w \in U} \left( \frac{P_v \cdot P_w}{10^6 \cdot D_{vw}^2} \cdot f_{vwij} \right) \right)$$

$$INTLCL_j = \sum_{\{v \in U, v \neq w\}} \left( \sum_{w \in U} \left( \frac{P_v \cdot P_w}{10^6 \cdot D_{vw}^2} \cdot g_{vwj} \right) \right)$$

where

$INT_{ij}$  = the through interaction score for entry external station  $i$  and survey external station  $j$

$INTLCL_j$  = the local interaction score for survey external station  $j$

$U$  = the set of each U.S. Census Bureau urban area and urban cluster which has its centroid within the study area; or has its centroid within 50 miles of the study area boundary; or has a population of at least 50,000 people and has its centroid within 250 miles of the urban area boundary

$v, w$  = indices for the urban areas in the set of urban areas  $U$

$P_v, P_w$  = the populations of  $v$  and  $w$

$D_{vw}$  = the non-congested least time route distance in miles from the centroid of  $v$  to the centroid of  $w$

$f_{vwij}$  = a binary variable which is 1 if the non-congested least time route from the centroid of  $v$  to the centroid of  $w$  passes through external stations  $i$  and  $j$ , and if the route segment between  $i$  and  $j$  is valid. The route segment is considered valid if (1) it passes through  $i$  before  $j$ ; and (2) it passes through the study area; and (3) it crosses the study area boundary only at  $i$  and  $j$ . Otherwise, the variable is 0.

**Table 6 continued**


---

$g_{vwj}$  = a binary variable which is 1 if the non-congested least time route from the centroid of  $v$  to the centroid of  $w$  passes through external station  $j$ , and if the centroid of  $v$  is inside the study area; and if the route segment between  $v$  and  $j$  is valid. The route segment is considered valid if it crosses the study area boundary only at  $j$ . Otherwise, the variable is 0.

---

### **Discussion of Some Candidate Predictor Variables**

For most of the candidate predictor variables, the meaning and importance of the variable is obvious from the variable definition. However, the interaction score variables, the route validity variable, and the average turns variable are complicated and warrant further discussion to make their meaning and importance more obvious. The “is commercial vehicle” variable is also discussed here, since it is an important new variable that has not been included in previous research.

#### *Interaction Score Variables*

The interaction score generates and distributes relative amounts of trips using a simple gravity model, assigns the trips to the roadway network, then checks to see if the trips pass through the study area. The gravity model assumes that all trips originate and terminate at the centroid of urban areas, and that the relative amount of travel between two urban areas is proportional to the product of the urban areas' populations, and inversely proportional to the square of the distance on the least time route between the urban area centroids. The interaction score is based on the work of Zipf (1946).

The results from the gravity model are assigned to the least-time route between urban area centroids, then these routes are checked to determine if they enter or exit the study area, or both, and to determine which external stations they use. Then the gravity model results are assigned to the appropriate external stations or external station pairs.

The interaction score, and the variables that are based on it, take into account the geographical distribution of land uses that generate and attract significant numbers of

trips, and the configuration of the roadway network that connects the land uses. These predictors are the basis of most travel demand models, and are two of the most important predictors of travel demand.

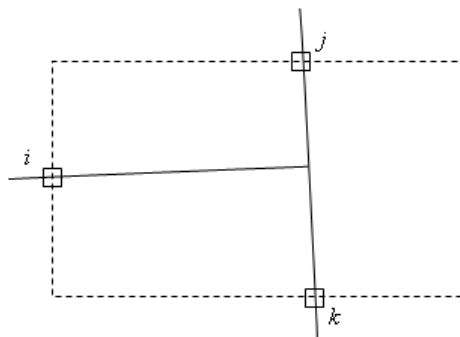
#### *Route Validity Variable*

For some external stations, the least-time route from one or more of the urban areas inside the study area to the external station passes through another external station. In this case the route is not valid. If none of the routes from internal urban area centroids are valid, then it is less likely that trips passing through the external station are E-I/I-E trips, and more likely that they are through trips. The variable  $RTELCL_j$  reflects this observation. It is 1 when at least one of the routes is valid, and is 0 when no routes are valid.

The route variable is illustrated by an example in Figure 2, where the dashed line represents a study area boundary, the solid lines represent roads, the small squares represent external stations, and the small circles represent urban areas. The least time routes from the urban area  $v$  to external stations  $i$  and  $j$  are valid. The least time routes from the urban area  $v$  to external stations  $k$  and  $l$  are not valid, since both routes pass through external station  $j$ . Therefore,  $k$  and  $l$  are less likely to have E-I/I-E trips and more likely to have through trips.



The average turns variable is illustrated in Figure 3. At least some of the through trips exiting the study area at external stations  $j$  and  $k$  can do so with having made any turns within the study area. Through trips exiting at external station  $i$  must have made one turn inside the study area. External stations  $j$  and  $k$  offer more directness to through trips, and the value of the average turns variable is lower for these external stations than for external station  $i$ , which offers less directness to through trips.



**Figure 3. Average turns example.**

#### *“Is Commercial Vehicle” Variable*

The variable *ISCV* is new in this research. None of the previous research made separate predictions for commercial vehicles and non-commercial vehicles. However, such a separation is important, since commercial vehicles likely have different through trip patterns. Commercial vehicles may have longer trips than non-commercial vehicles (see, for example, a Denver, Colorado travel survey which showed that the average trip duration for commercial vehicles and non-commercial vehicles is 32 minutes and 17 minutes respectively [Denver Regional Council of Governments, and Parsons Transportation Group, Inc 2000; Veras and Patil 2005]), so a greater proportion of commercial vehicle trips may be through trips than of non-commercial vehicle trips. In addition, since most commercial vehicles are also large vehicles, they place different

pavement, traffic flow, and air quality demands on the transportation system than do non-commercial vehicles.

### **Variables Not Selected as Candidate Predictor Variables**

Table 1 and Table 2 in the Literature Review show which variables were included in the final models of previous researchers. The set of candidate variables for this research included many of those variables, but not all of them. This section explains why some of the variables were not included.

From Table 1 the variables functional classification, percent pickups and vans, and marginal highway route were used in the final version of at least one previous model, but are not considered as candidate variables for this research.

The variable functional classification was not included for a number of reasons. First, functional classification tends to be somewhat subjective, especially for areas that transition from urban to rural. Since most external stations are in these types of areas, a functional classification variable may lead to inconsistencies in model development and application.

Second, several of the candidate predictor variables provide information that is very similar to that which would be provided by a functional classification variable. The roadway variables  $LANES_j$ ,  $DIVIDED_j$ , and  $LIMITED_j$  along with the traffic variables probably provide more than enough information to make up for the absence of the functional classification variable.

Third, previous research has not proven conclusively that functional classification is always a good predictor of through trips. The Modlin 1974 stage one model included functional classification, but it is unclear whether this model was compared with a model that did not include functional classification (Modlin 1974). Even Modlin himself did not include functional classification in his 1982 stage one model (Modlin 1982). The Han 2008 stage one model includes functional classification, but only as a single dummy variable indicating whether the road is a collector or local road or not (Han 2007; Han and Stone 2008).

The percent pickups and vans was not included as a candidate predictor variable because previous research did not show that it was always a good predictor. It was included in the the Modlin 1974 model, but Modlin did not include it in his 1982 model (Modlin 1974; Modlin 1982). In addition, this data was not available for some of the study areas, because pickup trucks and vans were aggregated with smaller cars and motorcycles.

Marginal highway route is a variable in the Han 2008 model which indicates whether an external station is on a highway route which cuts through the corner of a study area or almost parallels the study area boundary to create two external stations very close together on the same highway (Han 2007; Han and Stone 2008). Marginal highway route was not included as a candidate predictor variable for this research because several of the other candidate variables, such as the interaction score variables,  $RTELCL_j$ , and  $AVGTRN_j$  provided the same information in a less subjective way.

From Table 2, the variables “nearby major city”, “route continuity”, and “internal-external factor” were included in at least one previous model, but were not candidate variables for this research. “Internal- external factor” is not included because it is only necessary for combined models, and this research created a two-stage model. The variables “nearby major city” and “route continuity” were not included because the interaction score variables and  $AVGTRN_j$  provide the same information in a more comprehensive and less subjective way.

### **Data Sources for Predictor Variables**

Data for the traffic characteristics variables comes from pneumatic tube vehicle classification counts conducted with the external station surveys. Roadway data comes from Google Earth, which provides satellite images from several different years, so that the year of the image used and the year of the survey never differ by more than a few years. The study area characteristics come from data provided by the U.S. Census Bureau, and the U.S. Bureau of Labor Statistics. The source for data for the average turns variable is explained in the next chapter.



The interaction score variables depend on the location and population of urban areas, and on the least time routes between urban areas. As stated in the interaction score definition, the data for urban area locations and populations is provided by the U.S. Census Bureau, which publishes population estimates for each Census urban area and urban cluster, as well as provides a GIS file with polygons for all urban areas and clusters throughout the United States.

Least time routes between urban area centroids are extracted from the Bing Maps web service using a MS Visual Basic 2008 utility. The utility requires as input a list of the geographic coordinates of the study area external stations, as well as a list of the coordinates of the centroids of the urban areas included in the analysis. Using these coordinates, the utility submits requests for routes, and the Bing Maps web service returns route objects, which include a series of coordinates describing the shape and location of the route, the distance and travel time of the route, the route itinerary (directions) and other information about the route. For the interaction score variables, the utility requests routes between pairs of urban areas, then checks the route to determine if it passes through any of the external stations. For the local route validity variable, the utility requests routes from urban areas inside the study area to each external station, then checks the route to determine if it passes through any other external stations.

### **PRELIMINARY VARIABLE SELECTION**

The selection of candidate variables was based on the work of previous researchers, and on new theories about what variables have good potential for predicting through trips. From this set of candidate predictor variables, a new selection of variables was made based on forward selection, which is a more rigorous variable selection technique.

Forward selection begins with a model containing only a constant. Then, the one variable which has the lowest  $p$ -value in a likelihood ratio test when added to the constant only model is added to create a new model with one variable and the constant. Then, the one variable which has the lowest  $p$ -value when added to the one-variable model is added to create a new model with two variables and the constant. The forward

selection process continues in this manner, with variables added one at a time according to the results from a likelihood ratio test (Hosmer and Lemeshow 2000).

Usually, the forward selection process continues until no variable can be added with a  $p$ -value smaller than some pre-specified value, such as 0.05 or 0.01. However, preliminary analysis showed that following such a rule would result in selecting most of the candidate variables, and such a large model is not desirable for multiple reasons. First, selecting too many variables can result in a model which over-fits the data, meaning that the model fits noise in the data rather than the true pattern. Second, a large model would be more difficult to understand and interpret than a smaller model. Third, a large model would have higher data collection costs than a smaller model.

To help limit the number of variables selected, the model development process used the Akaike information criterion ( $AIC$ ), the Bayesian information criterion ( $BIC$ ) and adjusted rho-square ( $\bar{\rho}_C^2$ ). Each of these criteria is a measure of the log-likelihood of the model, penalized for the number of variables in the model. Lower values of  $AIC$  and  $BIC$ , and higher values of  $\bar{\rho}_C^2$  indicate a better model.

The  $AIC$  is given by (Koppelman and Bhat 2006)

$$AIC = \frac{-2(LL(\hat{\beta}) - K)}{M}$$

where

$AIC$  = Akaike information criterion;

$LL(\hat{\beta})$  = the log-likelihood for the estimated model;

$K$  = the number of parameters in the estimated model; and

$M$  = the number of covariate patterns in the sample, where a covariate pattern is a unique combination of the values of the predictor variables.

The  $BIC$  is given by (Koppelman and Bhat 2006)

$$BIC = \frac{-2(LL(\hat{\beta}) - K \cdot \log(K))}{M}$$

where

$BIC$  = Bayesian information criterion.

The adjusted rho square is given by (Greene)

$$\bar{\rho}_C^2 = 1 - \frac{LL(\hat{\beta}) - K}{LL(C) - K_{MS}}$$

where

$\bar{\rho}_C^2$  = adjusted rho squared with respect to the constants only model;

$LL(C)$  = the log-likelihood for the constant only model; and

$K_{MS}$  = the number of parameters in the constants only model (here equal to 1).

Normally, the model from forward selection with the best value of a criterion would be chosen. However, preliminary analysis showed that following this rule would also choose most of the predictor variables. Rather than the absolute value of each criterion, the rate of change of each criterion was used as a guide for forward selection. With this rule, a significant decrease in the rate of improvement of the criteria would suggest ending forward selection.

In addition to  $AIC$ ,  $BIC$  and  $\bar{\rho}_C^2$ , root mean square error ( $RMSE$ ) was used as a guide for forward selection, where a lower value indicates a better model. Although  $RMSE$  is not as statistically valid as the other three criteria, it does give a practical and intuitive sense of how well a model fits. The  $RMSE$  is given by

$$RMSE = \left( \frac{\sum_{m=1, \dots, M} (\hat{\pi}_m - p_m)^2}{M} \right)^{1/2}$$

where

$RMSE$  = the root mean square error;

$m$  = an index for each covariate pattern in the sample;

$p_m$  = the sample proportion through trips for covariate pattern  $m$ ; and

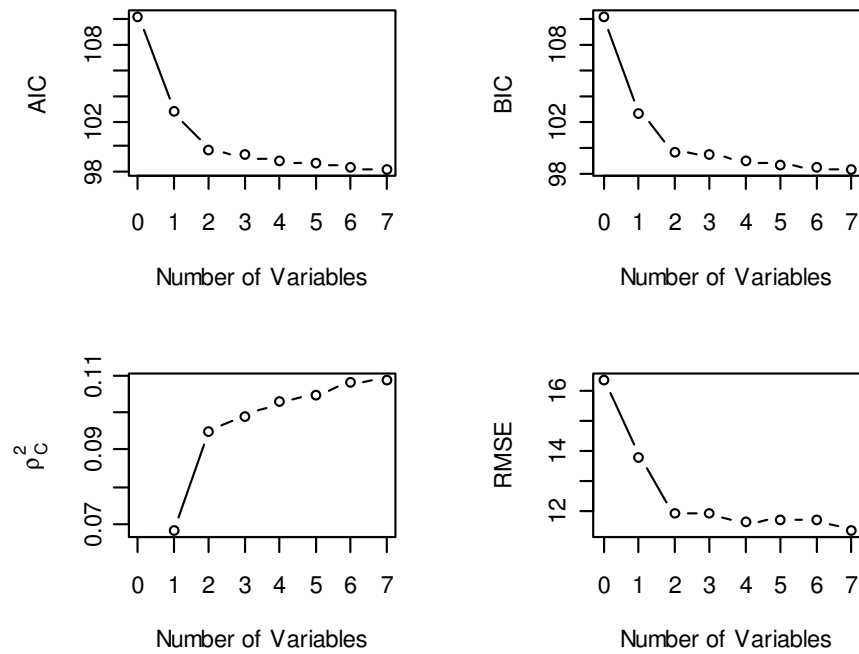
$\hat{\pi}_m$  = the proportion through trips as estimated by the model for covariate pattern  $m$ .

The results of the forward selection process are listed in Table 7 with the variables appearing in the order that they were added. Each row gives the values of the criteria for the model including the variables on that row and all previous rows. Each row also gives the  $p$ -value for a likelihood ratio test for the model with the variable on that row and all previous rows, compared to a model with only the variables on the previous rows.

**Table 7. Forward Selection Results for the Through Trip Split Model**

Variable	$AIC$	$BIC$	$\bar{\rho}_C^2$	$RMSE$	$P$
Constant	110.2	110.2		16.4	
$PINTTH_j$	102.7	102.7	0.068	13.8	$< 10^{-100}$
$ISCV_j$	99.7	99.7	0.095	11.9	$< 10^{-100}$
$LANES_j$	99.3	99.4	0.099	11.9	$10^{-38}$
$PROPLV_j$	98.8	98.9	0.103	11.6	$10^{-52}$
$EMP$	98.7	98.7	0.105	11.7	$10^{-15}$
$ADTALL$	98.3	98.4	0.108	11.7	$10^{-34}$
$RTELCL_j$	98.2	98.3	0.109	11.3	$10^{-16}$

The criteria as a function of the number of variables in the model are graphically presented in Figure 4. The plots show that each criterion improved quickly up to the second variable, where the rate of improvement of the criteria slowed significantly, suggesting that the model with two variables is the best model. However, to allow for the possibility that additional variables would be important to the model, variable selection continued.



**Figure 4. Forward selection criteria versus number of variables for the through trip split model.**

Variable selection stopped at the seventh variable, because the rate of improvement of the criteria continued to slow, and because all of the important variables had already been selected. Initial analysis showed that the last variable,  $RTELCL_j$  performed poorly, so it was dropped and the first six variables formed the preliminary model, called Model Ia. This model is presented in Table 8.

**Table 8. Model Ia**

Variable	Coeff.	Std. Err.	$z$	$p$	Mean of $x$
Constant	-3.10	$6.09 \times 10^{-2}$	$-5.09 \times 10^1$	$< 10^{-4}$	
$PINTTH_j$	1.92	$4.17 \times 10^{-2}$	$4.62 \times 10^1$	$< 10^{-4}$	$2.16 \times 10^{-1}$
$ISCV_j$	1.15	$3.04 \times 10^{-2}$	$3.80 \times 10^1$	$< 10^{-4}$	$1.25 \times 10^{-1}$
$LANES_j$	$-1.89 \times 10^{-1}$	$1.44 \times 10^{-2}$	$-1.31 \times 10^1$	$< 10^{-4}$	2.68
$PROPLV_j$	1.81	$1.42 \times 10^{-1}$	$1.27 \times 10^1$	$< 10^{-4}$	$1.55 \times 10^{-1}$
$EMP$	$-4.84 \times 10^{-7}$	$3.39 \times 10^{-8}$	$-1.43 \times 10^1$	$< 10^{-4}$	$6.82 \times 10^5$
$ADTALL$	$2.92 \times 10^{-6}$	$2.42 \times 10^{-7}$	$1.21 \times 10^1$	$< 10^{-4}$	$2.11 \times 10^5$
$\bar{\rho}_C^2 = 0.108$ , Number of responses = 74154					

## DIAGNOSTICS

Before continuing to the final variable selection, the model development process checks each observation using three model diagnostics,  $\Delta\chi_m^2$ ,  $\Delta D_m$ , and  $\Delta\beta_m$ . The first two diagnostics measure the effect of the observations with covariate pattern  $m$  on the model Pearson chi-square statistic and the model deviance, which are two summary measures of goodness of fit. The third diagnostic,  $\Delta\beta_m$ , detects covariate patterns whose observations have a large effect on the parameter estimates. Especially poor (high) values of these three diagnostics are useful in detecting covariate patterns which have data errors or whose observations are not fit well by the model. The three diagnostics are defined by (Hosmer and Lemeshow 2000)

$$\Delta\chi_m^2 = \frac{r_m^2}{(1-h_m)}$$

$$\Delta D_m = d_m^2 + \frac{r_m^2 h_m}{(1-h_m)}$$

$$\Delta\beta_m = \frac{r_m^2 h_m}{(1-h_m)^2}$$

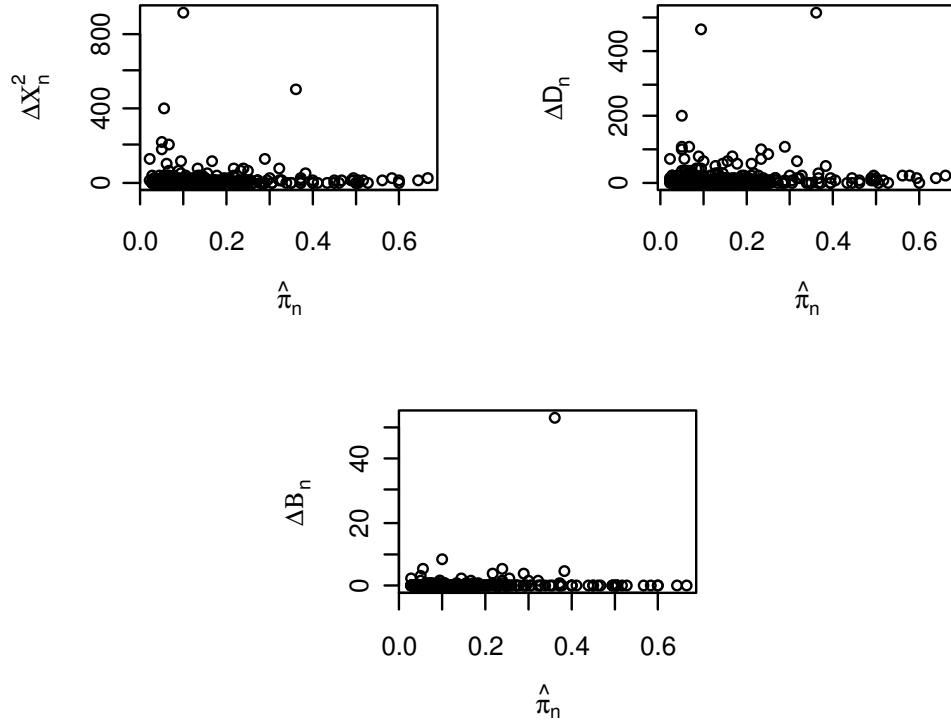
where

$h_m =$  the  $m^{th}$  diagonal element of the hat matrix as defined by Hosmer and Lemeshow (2000);

$r_m =$  is the Pearson residual as defined by Hosmer and Lemeshow (2000); and

$d_m =$  the deviance residual as defined by Hosmer and Lemeshow (2000).

Graphs of  $\Delta\chi_m^2$  and  $\Delta D_m$  versus  $\hat{\pi}_m$  are presented in Figure 5. Each of these graphs has 434 points: one commercial observation and one non-commercial observation for each of 217 external stations. Three points are especially high compared to the other points in each figure. These three points correspond to external stations 1213 in San Antonio, GR20 in Sherman-Denison, and 711 in Tyler, where in each case the model under predicted the proportion through trips. An investigation reveals no data errors (many errors have been identified and corrected in exploratory analysis), and shows that the observations are plausible. Attempts to find a variable which would improve the fit of the model for these points while not over-fitting the model were unsuccessful.



**Figure 5. Model Ia diagnostics.**

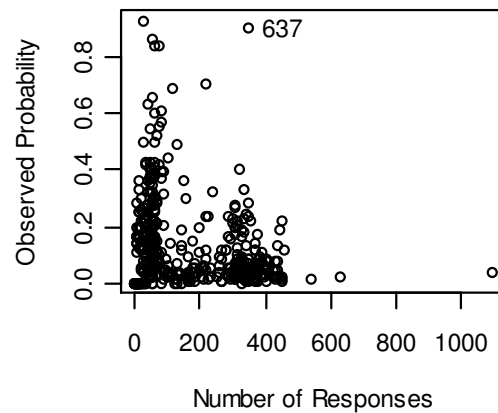
Figure 5 also includes a graph of  $\Delta\beta_m$  versus  $\hat{\pi}_m$ . One point, corresponding to observation number 637, is especially high compared to the other points. To further investigate its effect on the parameter estimates, observation 637 was removed from the data and Model Ia was fit again. The new parameter estimates are compared to the original parameter estimates in Table 9.



**Table 9. Model Ia with  
Observation 637 Removed**

Variable	Coefficient	Percent change
Constant	-3.01	-3
$PINTTH_j$	1.73	-10
$ISCV_j$	1.19	3
$LANES_j$	$-1.60 \times 10^{-1}$	-15
$PROPLV_j$	1.46	-19
$EMP$	$-4.81 \times 10^{-7}$	-1
$ADTALL$	$2.53 \times 10^{-6}$	-13
$\bar{\rho}_C^2 = 0.123$		
Number of responses = 73805		

Several of the parameter estimates changed significantly after removing observation 637 from the dataset, prompting a more thorough investigation of this observation. Observation 637 represents 349 individual responses from drivers of non-commercial vehicles exiting the San Antonio study area at external station 1213. Of these individual responses, about 91 percent were through trips, and the remaining 9 percent were E-I/I-E trips. When compared to the other observations in the data set, observation 637 has a fairly high number of responses, and a very high observed percent through trips (see Figure 6). The high percent through trips and high number of responses may cause observation 637 to have a large effect on the parameter estimates.



**Figure 6. Observed through trip probability versus number of responses.**

Although observation 637 has a high influence on the model parameters, no reason was found to remove it from the data set. The response and predictor variable data for observation 637, as well as the process for computing the original and new parameter estimates were examined, but no errors were found. In addition, looking at a map of the San Antonio study area and its external stations showed that observation 637 is very plausible. The external station is on a route that cuts through the corner of the study area, resulting in a pair of external stations that are very close together and that are on the same highway, which in turn results in many trips being through trips entering at one of the external stations at exiting at the other.

Retaining observation 637 in the dataset allows the model development process to produce more valid and usable results. If the model development process includes observation 637, then the model can be applied to new study areas without any need for excluding external stations with characteristics similar to those of the external station

corresponding to observation 637. In addition, with observation 637 in the data set, the theoretical conclusions drawn from the model development results are more general in nature.

## **FINAL VARIABLE SELECTION**

The results from the forward selection process suggested that a model with only two variables was the most appropriate model. However, to allow for the possibility that a richer model would actually be more appropriate, six variables were selected and retained throughout the model diagnostics process. Next the model development process investigated more thoroughly the hypothesis that the two-variable model is the better model.

The investigation was based on the following experiment: Take a random sample of study areas from the set of all 13 study areas, then fit a model using the results from the external surveys for the sampled study areas. Repeat this a number of times and then compare the parameter estimates from each repetition. Variables whose parameter estimates change relatively little between repetitions of the experiment are better predictors than variables whose parameter estimates change much.

The results from this approach roughly give some of the same information as the standard error in Table 8, since both measure the variability of the parameter estimates. However, this approach has the advantage that it measures parameter estimate variability by sampling whole study areas at a time, which gives confidence that the results can be extended to an entirely new study area.

To carry out the experiment, the set of 13 study areas was randomly divided into four groups, as presented in Table 10. For each new model, one group of study areas was removed from the dataset, and a model was fit to the remaining data, to produce four sets of parameter estimates. Each new parameter estimate as a relative change from the original parameter estimates is presented in Table 11. The first two variables have significantly smaller changes than do the last four. The largest change in the first two variables is 16 percent, while the last four variables change by at least 29 percent at least once, and three of the last four change by at least 51 percent at least once.

**Table 10. Study Area Groups**

Group 1	Amarillo, San Antonio and Waco
Group 2	Austin, Lubbock and Sherman-Denison
Group 3	Dallas-Fort Worth, Longview, San Angelo and Wichita Falls
Group 4	Abilene, Midland-Odessa and Tyler

**Table 11. Model Ia Relative Parameter Estimate Changes (percent)**

Variable	Group Removed			
	1	2	3	4
Constant	-7	21	-1	-9
$PINTTH_j$	-12	1	7	4
$ISCV_j$	16	-6	-9	-2
$LANES_j$	9	-51	17	19
$PROPLV_j$	-23	22	29	-16
$EMP$	4	23	-56	-14
$ADTALL$	-14	59	-20	-29

The results from this investigation confirmed the hypothesis that the smaller two variable model may be the better model. One of these two variables is  $PINTTH_j$ , which is the interaction score for through trips as a portion of the interaction score for all trips. However, some of the external stations have no interaction scores at all. For these external stations the portion is not defined, and the variable is set to be zero. Thus, the meaning of a zero value for this variable is ambiguous, because it could mean that the external station has no interactions at all, or it could mean that the external station has interactions, but none of them are interactions for through trips. To allow the model to distinguish between these two situations, the variable  $INTTLI_j$  was added to the model. This variable is 0 when the external station has no interaction scores, and is 1 when the external station has any interaction score. Thus it serves as an adjustment to the model to

distinguish between the two cases when  $PINTTH_j$  is zero. The resulting model is Model Ib, and is presented in Table 12.

**Table 12. Model Ib**

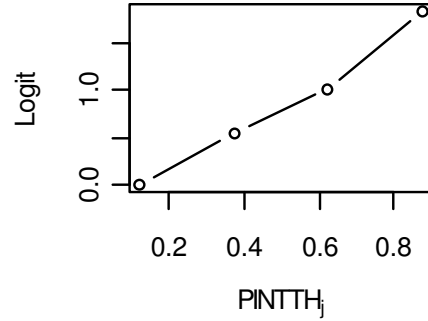
Variable	Coeff.	Std. Err.	$z$	$p$	Mean of $x$
Constant	-2.94	$2.92 \times 10^{-2}$	$-1.01 \times 10^2$	$< 10^{-4}$	
$PINTTH_j$	2.38	$4.66 \times 10^{-2}$	$5.10 \times 10^1$	$< 10^{-4}$	$2.16 \times 10^{-1}$
$ISCV_j$	1.16	$3.01 \times 10^{-2}$	$3.84 \times 10^1$	$< 10^{-4}$	$1.25 \times 10^{-1}$
$INTTL_j$	$-2.35 \times 10^{-1}$	$3.83 \times 10^{-2}$	-6.13	$< 10^{-4}$	$6.97 \times 10^{-1}$
$\bar{\rho}_C^2 = 0.096$ , Number of responses = 74154					

## MODEL REFINEMENT

To this point, the model development assumed that the continuous variables are linear in the logit, and that the effect of each variable does not vary across the levels of any of the other variables. The model development process next tested each of these assumptions.

To test the first assumption the model development process uses the logit step test, which is performed as follows. First, divide the continuous variable into four groups of equal intervals, or divide it into four groups based on the quartiles of the variable. Then recode the continuous variable as a categorical variable with a set of three design variables. The design variables correspond to the second, third, and fourth groups of the continuous variable, and the first group acts as the base class. Next, fit a model, replacing the continuous variable with the new categorical variable. Finally, plot the parameter estimate for each design variable against the midpoint of the corresponding group of the continuous variable. For the first group, plot zero against its midpoint. If the continuous variable is linear in the logit, then the plotted points should show a linear relationship. If not, then the shape of the plot can suggest possible transformations of the continuous variable to make it linear in the logit (Hosmer and Lemeshow 2000). The plot for the step test of  $PINTTH_j$ , the only continuous variable in the model, is presented

in Figure 7. The plot does not show evidence that  $PINTTH_j$  is not linear in the logit, so the variable was not transformed.



**Figure 7. Step test for  $PINTTH_j$ .**

To test the second assumption, that the effect of each variable does not vary across the levels of any other, the model development process compares Model Ib to a model with added interactions. With three variables in Model Ib, three two-variable interactions are possible, but only the interactions with  $ISCV$  were tested, defined by

$$SIPINTTH_j = ISCV \cdot PINTTH_j$$

and

$$SIINTTLI_j = ISCV \cdot INTTLI_j.$$

The two interactions were added to Model Ib to create Model Ic, which is presented in Table 13. The interactions have  $p$  values that are much higher than the original parameter estimates. Therefore, Model Ic was not retained, and Model Ib is the final through trip split model.

**Table 13. Model Ic**

Variable	Coeff.	Std. Err.	$z$	$p$	Mean of $x$
Constant	-2.97	$3.26 \times 10^{-2}$	$-9.10 \times 10^1$	$< 10^{-4}$	
$PINTTH_j$	2.43	$5.29 \times 10^{-2}$	$4.59 \times 10^1$	$< 10^{-4}$	$2.16 \times 10^{-1}$
$ISCV_j$	1.30	$6.68 \times 10^{-2}$	$1.95 \times 10^1$	$< 10^{-4}$	$1.25 \times 10^{-1}$
$INTTL_j$	$-2.18 \times 10^{-1}$	$4.43 \times 10^{-2}$	-4.91	$< 10^{-4}$	$6.97 \times 10^{-1}$
$SIPINTTH_j$	$-2.56 \times 10^{-1}$	$1.11 \times 10^{-1}$	-2.31	0.0212	$3.16 \times 10^{-2}$
$SIINTTL_j$	$-7.49 \times 10^{-2}$	$8.83 \times 10^{-2}$	$-8.49 \times 10^{-1}$	0.3961	$9.57 \times 10^{-2}$
$\bar{\rho}_C^2 = 0.096$ , Number of responses = 74154					

## MODEL EVALUATION

With the final model chosen, attention turned to evaluating its performance. The goal of this research is to develop a model that can be applied with reasonably accurate results to Texas study areas, including study areas that are not in the dataset for this research. The model evaluation simulates applying the model to new study areas using cross validation, which fits the model using a randomly selected segment of the available data, then tests the model on the remaining data. The model fitting and testing was repeated four times, using each of the four groups as defined in Table 10 as the test datasets, and the remaining data in each case as the model fitting dataset.

The model evaluation process focused on comparing the observed through trip splits to the through trip splits predicted by the model. To facilitate this comparison, a cross-classification table was created for each of the four test cases, where each observation was classified into ranges of observed and predicted through trip splits (see Table 14 through Table 17). Ideally, all observations would fall in the cells on the main diagonal of each table, which would indicate that the predictions are close to the observations. In reality many of the observations are not on the main diagonal. However,

large deviations from the main diagonal are infrequent, indicating that the model performs fairly well.

**Table 14. Predicted versus Observed Percent Through Trips for Group 1**

Observed(%)	Predicted (%)						
	0-5	5-10	10-20	20-40	40-60	60-80	80-100
0-5	19	3	15	0	0	0	0
5-10	6	4	5	0	0	0	0
10-20	1	4	8	7	0	0	0
20-40	1	1	8	6	3	0	0
40-60	0	0	2	1	2	0	0
60-80	0	0	0	0	0	0	0
80-100	0	0	0	1	1	0	0

**Table 15. Predicted versus Observed Percent Through Trips for Group 2**

Observed(%)	Predicted (%)						
	0-5	5-10	10-20	20-40	40-60	60-80	80-100
0-5	16	7	6	0	0	0	0
5-10	7	3	6	1	0	0	0
10-20	4	3	14	2	0	0	0
20-40	2	0	11	4	0	0	0
40-60	0	0	2	2	0	0	0
60-80	0	0	1	0	1	0	0
80-100	0	0	0	0	0	0	0



**Table 16. Predicted versus Observed Percent Through Trips for Group 3**

Observed(%)	Predicted (%)						
	0-5	5-10	10-20	20-40	40-60	60-80	80-100
0-5	12	30	20	1	0	0	0
5-10	3	7	15	3	0	0	0
10-20	1	4	17	2	0	0	0
20-40	0	1	19	13	6	0	0
40-60	0	0	2	1	2	1	0
60-80	0	0	0	2	1	0	0
80-100	0	0	0	0	0	1	0

**Table 17. Predicted versus Observed Percent Through Trips for Group 4**

Observed(%)	Predicted (%)						
	0-5	5-10	10-20	20-40	40-60	60-80	80-100
0-5	5	18	12	1	0	0	0
5-10	0	6	5	2	0	0	0
10-20	0	1	11	2	1	0	0
20-40	0	1	5	3	1	0	0
40-60	0	0	0	3	0	0	0
60-80	0	1	0	0	0	0	0
80-100	0	0	0	1	1	0	0

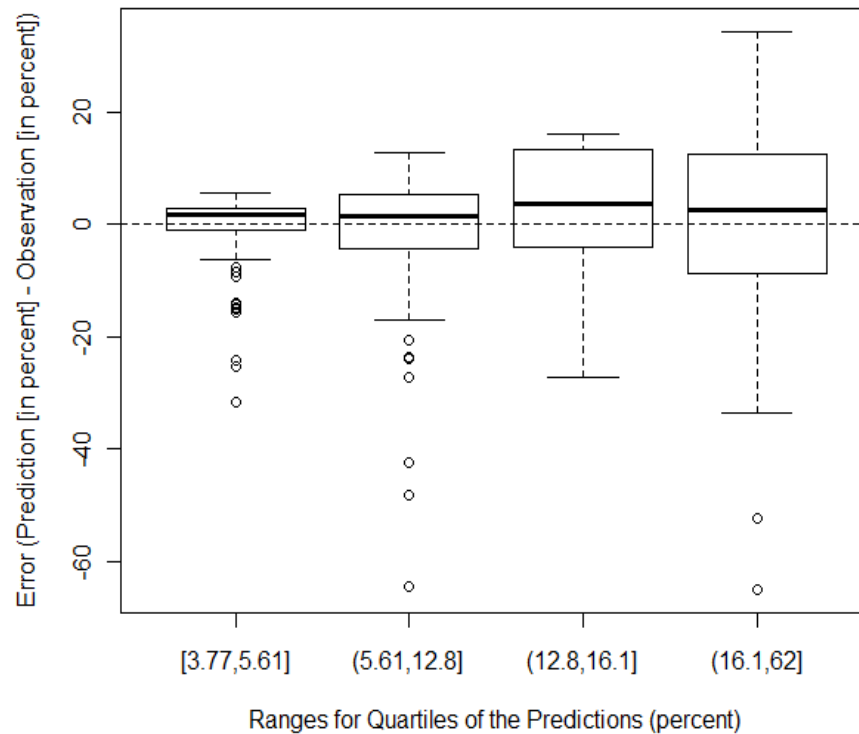
To allow further analysis of the model performance, the prediction error was calculated and investigated. The prediction error is defined to be the predicted through trip split (in percent) minus the observed through trip split (in percent). Positive values of the error indicate that the prediction is greater than the observation, and negative values of the error indicate that the prediction is less than the observation. The investigation of the prediction error used the results from all four test cases combined into a single dataset.

The prediction error was investigated by examining the distribution of the absolute values of the error. Of all the predictions, 49 percent had an absolute error of less than 5 percent, 19 percent of the observations had an absolute error between 5 and 10 percent, and 22 percent of the observations had an error between 10 and 20 percent. Only 10 percent of the observations had an absolute error greater than 20 percent (see Table 18), indicating that the model performs well.

**Table 18. Summary of Through  
Trip Split Model Prediction Errors**

Range of absolute error(%)	Percent of observations
0-5	49
5-10	19
10-20	22
20-40	9
40-100	1

To further investigate the distribution of the errors, the data set was ordered by the predicted through trip split, then separated into four intervals with very nearly the same number of observations in each. Then, a box plot of the error was made for each of the four intervals. The bottom and the top of each box respectively represent the 25<sup>th</sup> percentile and 75<sup>th</sup> percentile of the error distribution. The heavy line inside each box represents the median of the error. Each “whisker” either extends to the most extreme prediction error, or is as long as 1.5 times the difference between the 25<sup>th</sup> and 75<sup>th</sup> percentiles. Any prediction errors falling outside this range are plotted as small circles. The box plots are presented in Figure 8.



**Figure 8. Box plots of the split model prediction errors for each quartile of the prediction values.**

The box plots indicate that the predictions fit the observations fairly well. The center of the distribution suggested by each box plot is close to zero, and in general small errors are more likely than more extreme errors. The box plots also show that prediction errors increase as the value of the prediction becomes larger, indicating that the model performs better when predicting low through trip splits than when predicting high through trip splits.

## RESULTS

This chapter described the model development process for the through trip split model, which estimates the portion of all trips that are through trips. The model evaluation showed that the predictions fit the observations fairly well, indicating that the split model can be useful for practical applications. Because the final model included the variable *ISCV* (which is a binary variable indicating whether the estimation is for commercial vehicles or for non-commercial vehicles) the through trip split model can be written with two equations: one for each vehicle type. The equation for commercial vehicles is

$$g(\mathbf{x}_j) = -2.94 + 2.38PINTTH_j + 1.16 - 0.235INTTLI_j$$

and the equation for non-commercial vehicles is

$$g(\mathbf{x}_j) = -2.94 + 2.38PINTTH_j - 0.235INTTLI_j$$

where

$j$  = an index for the external station for which the estimation is made;

$\mathbf{x}_j$  = the vector of predictor variables for external station  $j$ ;

$g(\mathbf{x}_j)$  = the utility of through trips for vehicles exiting the study area at external station  $j$ ;

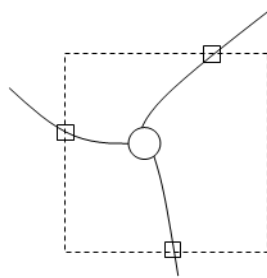
$PINTTH_j$  = the proportion through trips at the external station as predicted by the interaction score; and

$INTTLI_j$  = a binary variable which is 1 if the external station has any interaction scores, and is 0 if it has no interaction scores.

### Variables Not Included in the Final Model

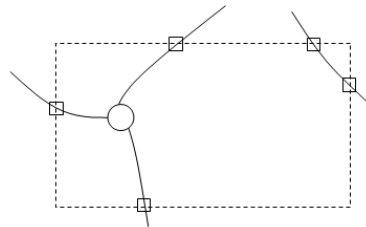
One of the most important observations about the through trip split model is that it does not include variables for average daily traffic, percent trucks or the study area population, whereas all four of the final stage one models from previous research do. The key to understanding this apparent discrepancy lies in the fact that the value of the information provided by these variables, along with all the traffic, roadway and study area variables, depends heavily on the location of the study area boundary.

Consider the study area in Figure 9, where the dotted line represents the study area boundary, the solid lines represent roads, the small circle represents an urban area, and the small squares represent external stations. For this study area, a reasonable assumption is that increasing the population of the urban area would also increase the number of trip ends inside the study area, and lead to a decrease in proportion through trips at the external stations.



**Figure 9. A well-behaved study area.**

Now consider Figure 10, where the study area has been extended, and now includes one additional highway and two additional external stations. For this extended study area, an increase in the urban area's population would probably also lead to an increase in the number of trip ends inside the study area. However, this increase would have little effect on proportion through trips at the two new external stations, since these external stations are on a highway that does not access the urban area. Thus, depending on the location of the study area boundary, the population variable (and the other study area characteristics variables) may or may not be a good predictor of percent through trips at a particular external station.



**Figure 10. An ill-behaved study area.**

Figure 9 and Figure 10 and also help illustrate how the value of the traffic and roadway variables depends on the location of the study area boundary. For the study area in Figure 9, a reasonable assumption is that the external stations with higher average daily traffic also have a higher proportion through trips, since the trips at an external station with a high ADT probably have a high average trip length.

Now consider the two additional external stations in Figure 10. These two external stations are on a highway which does not access the urban area in the study area, so it is unlikely that a trip that enters or exits the study area at one of the external stations would begin or end inside the study area. These two external stations will probably have a high proportion of through trips, regardless of the ADT or average trip length at the external stations. Thus the value of ADT (and all the other traffic and roadway variables) for predicting proportion through trips depends on the location of the study area boundary.

The population, ADT, and percent trucks variables were important in previous research because most or all of the study areas used to develop the models are similar to the study area in Figure 9, in that all of the external stations are on highways that provide direct access to the central urban area of the study area. In Texas however, most of the study areas are defined so that the boundaries follow county lines. This partially arbitrary definition of study area boundaries leads to some instances of external stations

like the two additional ones in Figure 10; they are on highways that don't provide access to or otherwise have very little to do with the urban area or areas inside the study area.

### **Variables In the Final Model**

#### *“Is Commercial Vehicle”*

A unique aspect of the through trip split model is that it makes two separate predictions: one for each vehicle type. All previous models make a single prediction for all vehicle types. The separate predictions are very useful to the through trip modeling process. The fact that *ISCV* is one of the more important variables in the forward selection process indicates that commercial vehicles have different through trip patterns than do non-commercial vehicles. Making separate predictions allows the model to capture this difference. Commercial vehicles and non-commercial vehicles are also different in the demands they place on the transportation system, so knowing the through proportion of each type individually helps to have a more accurate understanding of transportation system needs.

The estimated coefficient for *ISCV* indicates that in general, a larger proportion of commercial vehicle trips than of non-commercial vehicle trips are through trips. This results from the fact that commercial vehicles have longer trips on average, so are more likely to travel completely through the study area.

At first, the *ISCV* variable may appear to be similar to the traffic and roadway variables, in that they all help describe the lengths of trips passing through an external station. So the question arises, why is *ISCV* an important variable, when the traffic and roadway variables are not. The answer comes from the fact that almost every covariate pattern in the data sample can be paired with another covariate pattern in the sample which is identical in every way except for the value of *ISCV*. Thus, the effect of *ISCV* can be determined because the effect of all the other variables is controlled. This is not the case for the traffic and roadway variables.

#### *Interaction Score Variables*

According to the forward selection process, the most important variable for predicting the proportion of through trips is *PINTTH<sub>j</sub>*, which is the proportion of through

trips as predicted by the interaction score (see Table 6). Unlike the traffic, roadway, and study area variables, it does not place the study area on an “island”, separated from its regional context. Rather, it uses detailed information about the location of the study area boundary and external stations in relation to the location of urban areas within and surrounding the study area, and the locations of routes connecting study areas. Because it uses this detailed information about the regional context of the study area, the value of  $PINTTH_j$  does not depend on the location of the study area boundary; in fact, the variable would even function well for a study area of random size, shape, and location, as long as the study area boundary does not pass through any urban area.

That  $PINTTH_j$  is a good predictor should be no surprise, since it is based on the gravity model concept, which drives most travel demand models in use today.

The variable  $PINTTH_j$  is based on the interaction score, which simplifies the geographical distribution of trip ends by assuming that all trips begin and end at the centroid of an urban area. This assumption greatly reduces the computational demands of calculating the interaction score, but it also causes some external stations to have no interaction scores at all. For these external stations,  $PINTTH_j$  cannot be calculated, and it is set to zero. At the same time, however, the fact that an external station has no interaction scores may provide information on the proportion through trips. For this reason, the variable  $INTTLI_j$  is included in the final model.

At first, the negative coefficient for  $INTTLI_j$  seems to indicate that external station that have an interaction score have a smaller proportion through trips than external stations that don't have an interaction score. However, the relationship is not so simple, because the positive contribution from  $PINTTH_j$  can be (and is probably most often) greater than the negative contribution from  $INTTLI_j$ .



## **CHAPTER V**

### **THROUGH TRIP DISTRIBUTION MODEL DEVELOPMENT AND EVALUATION**

The through trip distribution model is the second model in the two-model system developed by this research. The first model predicts the portion of all trips at an external station that are through trips. The through trip distribution model then distributes the through trips by predicting the proportion of all through trips exiting the study area at external station  $j$  that entered the study area at each external station  $i$ . The development of the through trip distribution model followed the same general process as that of the through trip split model: the first step was to choose candidate predictor variables; then the preliminary model was formed by choosing variables from the set of candidate variables based on the results of forward selection. Then, a second and final variable selection was made from the variables in the preliminary model. Next, transformations of the selected variables were tested to form the final model. Finally, the performance of the final model was evaluated. This chapter describes each of these parts of the through trip distribution model development process.

#### **CANDIDATE PREDICTOR VARIABLES**

The first step of the model development process was to select a set of candidate predictor variables. The purpose of this step is to provide a set of variables that merit further analysis. The selection of some of the candidate variables was based on the results of previous research, but other candidate variables were new. This section defines each variable, discusses the new and more complicated variables in more detail, explains why some variables were not selected as candidate predictor variables, and gives the variable data sources.

#### **Variable Definitions**

Each of the candidate predictor variables for the through trip distribution model is presented in Table 19.

---

**Table 19. Candidate Predictor Variables for the Through Trip Distribution Model**  
**Traffic Characteristics**

---

$ADTALL_i =$  The average daily traffic (ADT) for external station  $i$  for all vehicle types, where ADT is average non-holiday weekday 24-hour two-way count of vehicles passing through the external station

$ADTLV_i =$  The ADT for external station  $i$  for large vehicles, where large vehicles are vehicles belonging to classes 4 through 13 of the Federal Highway Administration (FHWA) vehicle classification system

$$PROPLV_i = \frac{ADTLV_i}{ADTALL_i}$$

The portion of the total ADT that is large vehicles

$$PADTSV_{ij} = \frac{ADTSV_i}{\sum_{\{q \in E, q \neq j\}} ADTSV_q}$$

The small vehicle ADT as a portion of the total small vehicle ADT across all external stations in the study area except the survey external station, where  $ADTSV_i$  and  $ADTSV_q$  are the ADTs for small vehicles for external stations  $i$  and  $q$  respectively,  $j$  is the survey external station, and small vehicles are vehicles belonging to classes 1 through 3 of the FHWA vehicle classification system.

---

Table 19 continued

---


$$PADTLV_{ij} = \frac{ADTLV_i}{\sum_{\{q \in E, q \neq j\}} ADTLV_q}$$

The large vehicle ADT as a portion of the total large vehicle ADT across all external stations in the study area except the survey external station, where  $ADTLV_i$  and  $ADTLV_q$  are the ADTs for large vehicles for external stations  $i$  and  $q$  respectively, and  $j$  is the survey external station

$$PADTALL_{ij} = \frac{ADTALL_i}{\sum_{\{q \in E, q \neq j\}} ADTALL_q}$$

The ADT for all vehicles as a portion of the ADT for all vehicles across all external stations in the study area except the survey external station, where  $ADTALL_i$  and  $ADTALL_q$  are the ADTs for all vehicles for external stations  $i$  and  $q$  respectively, and  $j$  is the survey external station

---

**Table 19 continued**  
**Roadway Characteristics**

---

$LANES_i$  = Total number of lanes in both directions at external station  $i$ . For example, the value of  $LANES_i$  for an external station with two lanes in each direction would be 4. The lane count only includes main through lanes. Any turning lanes, median left turn lanes, climbing lanes or passing lanes are not counted.

$4LANE_i$  = A binary variable which is 1 when  $LANES_i$  is greater than or equal to 4, and is 0 otherwise

$DIVIDED_i$  = A binary variable which is 1 when, in the area of external station  $i$ :  
 (1) the two directions of traffic are separated by either a non-traversable barrier, such as a wall or railing, or by a non-paved area which is not intended for traffic, such as a grassy median; and (2) opportunities for left turns across the barrier or non-paved area at an intersection are less frequent than is typical for an urban arterial. The variable is 0 otherwise.

$LIMITED_i$  = A binary variable which is 1 when the roadway in the area of external station  $i$  is a limited-access facility, which means that access to the roadway is only provided by ramps. For areas where the roadway transitions from limited access to non-limited access the variable is 1. The variable is 0 otherwise.

---

Table 19 continued

## Measures Separation between External Station Pairs

- $DISTRO_{ij}$  = The great circle distance in miles divided by the non-congested least time route distance in miles from external station  $i$  to external station  $j$
- $RSPEED_{ij}$  = The great circle distance in miles divided by the non-congested least time route duration in hours from external station  $i$  to external station  $j$
- $TURNS_{ij}$  = The number of turns on the non-congested least time route from external station  $i$  to external station  $j$ , where a turn is any movement at an intersection besides the main through movement
- $RAMPS_{ij}$  = The number of freeway ramps, including on-ramps, off-ramps and freeway-to-freeway ramps, on the non-congested least time route from external station  $i$  to external station  $j$

## Interaction Score Variables

- $INT_{ij}$  = The interaction score for external stations  $i$  and  $j$  as defined in Table 6
- $PINT_{ij} = \frac{INT_{ij}}{\sum_{\{q \in E, q \neq j\}} INT_{qj}}$
- The proportion of the through interaction score at external station  $j$  that originates from external station  $i$ . If  $\sum_{\{q \in E, q \neq j\}} INT_{qj} = 0$ , then  $PINT_{ij} = 0$
- $PINTI_{ij}$  = A binary variable which is 1 if  $PINT_{ij}$  is greater than 0, and is 0 otherwise

**Table 19 continued**  
**Route Validity**

$ROUTE_{ij}$  = A binary variable which is 1 if the least time route from external station  $i$  to external station  $j$  is valid, and is 0 otherwise. The route is valid if (1) it passes through the study area, and (2) it crosses the study area boundary only at external stations  $i$  and  $j$ .

---

**Results from the Through Trip Split Model**

---

$CVSPLT_i$  = Proportion through trips at external station  $i$  for commercial vehicles as predicted by the through trip split model

$NCSPLT_i$  = Proportion through trips at external station  $i$  for non-commercial vehicles as predicted by the through trip split model

$ADTTHR_i = CVSPLT_i \cdot ADTLV_i + NCSPLT_i \cdot ADTSV_i$

The total through volume at external station  $i$  as predicted by the through trip split model

---

### **Discussion of Some Candidate Predictor Variables**

The meaning and importance of most of the candidate predictor variables is clear from the variable definitions. However, some of the variables are more complicated, or are new variables that have not been used by other researchers. These variables are discussed here in greater detail, with the exception of the variables based on the interaction score, which is discussed in the previous chapter.

#### *Measures of Separation between External Stations*

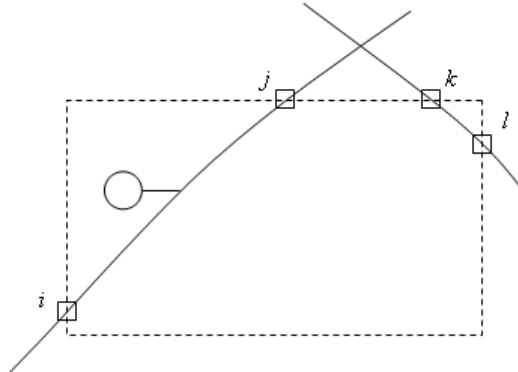
The purpose of the measures of separation between external station pairs is to quantify the likelihood that a trip passing through one external station would also pass through a second external station. The approach is to measure the directness of travel between external station pairs, with the assumption that external station pairs with more direct connecting routes are more likely to share trips than external station pairs with less

direct connecting routes.  $DISTRO_{ij}$  and  $RSPEED_{ij}$  measure the distance and time separation between external stations, while normalizing for the great circle distance between external stations. High values of these two variables indicate high directness.  $TURNS_{ij}$  and  $RAMPS_{ij}$  measure the separation between external stations using the number of turns and ramps. High values of these variables indicate low directness.

#### *Route Validity Variable*

For some pairs of external stations, the least time route connecting the two passes through one or more other external stations, or the route does not pass through the study area. In this case the route is not valid. It is unlikely two external stations that do not have a valid connecting route would exchange through trips.

The route validity variable is illustrated in Figure 11, where the dotted line represents the study area boundary, the solid lines represent roads, and the small squares represent external stations. If external station  $i$  is the survey external station, then the route from external station  $j$  is valid, but the routes from external stations  $k$  and  $l$  are not valid. Therefore,  $i$  probably exchanges its through trips mostly with external station  $j$ .



**Figure 11. Through route validity example.**

### **Variables Not Selected as Candidate Predictor Variables**

Of the variables that were included in previous researchers' combined and stage two models (see Table 2 and Table 3) most are included as candidate variables for this research. However, some are not candidate variables. This section explains why they are not included.

The variables functional classification, route continuity, Zipf's probability factor, and Horowitz's weight were included in at least one previous model, but are not candidate predictor variables for this research. Functional classification is not included here for the same reasons it was not included as a candidate variable for the through trip split model. The variable route continuity is not included because the external station separation variables give the same information in a more comprehensive way. Zipf's probability factor and Horowitz's weight are replaced by the interaction score variables.

Unlike the through trip split model, this model does not include *ISCV* as a candidate predictor variable. Preliminary analysis showed that many of the survey external stations had very few through observations. Making separate predictions for commercial and non-commercial vehicles would result in many instances where the through trip distribution estimation is based on only one or two observed through trips. Combining the two vehicle types allows the estimation to be based on more through trips.

### **Data Sources for Predictor Variables**

Data for the traffic characteristics variables comes from pneumatic tube vehicle classification counts conducted with the external surveys. Roadway data comes from Google Earth, which provides satellite images from several different years, so that the year of the image used and the year of the survey never differ by more than a few years.

The interaction score variables depend on the location and population of urban areas, and on the least time routes between urban areas. As stated in the interaction score definition, the data for urban area locations and populations is provided by the U.S. Census Bureau, which publishes population estimates for each Census urban area and



urban cluster, as well as provides a GIS file with polygons for all urban areas and clusters throughout the United States.

Least time routes between urban area centroids are extracted from the Bing Maps web service using a MS Visual Basic 2008 utility. After extracting the routes, the utility analyzes them to determine which external stations the route passes through (if any) and if the route segments are valid, as described in the interaction score definition.

The same utility is also used to extract routes for the route validity variable, and for the external station separation variables.

### PRELIMINARY VARIABLE SELECTION

Forward selection as described in the previous chapter was used to select variables from the set of candidate predictor variables, with two changes. First, the criterion  $\bar{\rho}_c^2$  was replaced by the criterion  $\bar{\rho}_0^2$ . Both of these criteria measure the log-likelihood of the model, while penalizing larger models. The first criteria,  $\bar{\rho}_c^2$ , measures the log-likelihood of the model with respect to a model with only a constant. For the through trip distribution model, a model with only a constant would not be appropriate, since all alternatives have the same utility function. The variable  $\bar{\rho}_0^2$  measures the log-likelihood with respect to a model which gives equal likelihood to each alternative. The adjusted rho squared with respect to the equal likelihood model is given by (Koppelman and Bhat 2006):

$$\bar{\rho}_0^2 = 1 - \frac{LL(\hat{\beta}) - K}{LL(0)}$$

where

$\bar{\rho}_0^2$  = adjusted rho squared with respect to the equal likelihood model;

$LL(\hat{\beta})$  = the log-likelihood of the estimated model;

$LL(0)$  = the log-likelihood for the model assigning equal likelihood to all alternatives; and

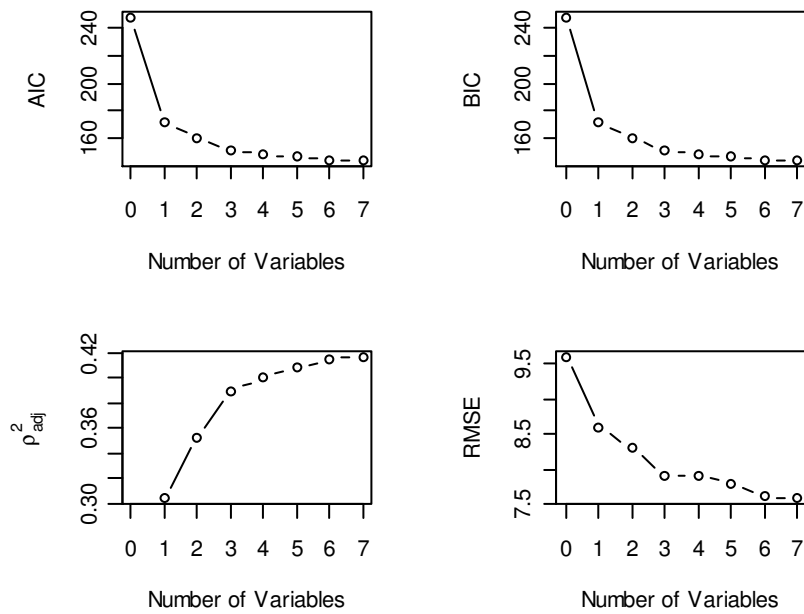
$K$  = the number of parameters in the estimated model.

The second change was that the forward selection process starts with the equal-likelihood model, rather than the constant only model.

The results from the forward selection process are presented in Table 20. The row labeled “None” presents the results for the equal-likelihood model, and each following row give the variables in the order they were added to the model. The results are presented graphically in Figure 12, where the points corresponding to 0 variables are for the equal-likelihood model. As expected, each of the criteria improves significantly from the equal-likelihood model to the model with 1 variable. For the models with one or more variables, the rate of improvement of the criteria slows after the model with 3 variables. Forward selection stopped at seven variables, because the rate of improvement of the criteria continued to slow, and because the most important variables had been selected.

**Table 20. Forward Selection Results for the Through Trip Distribution Model**

Variable	<i>AIC</i>	<i>BIC</i>	$\bar{\rho}_C^2$	<i>RMSE</i>	<i>P</i>
None	247.1	247.1		9.64	
<i>PINTI<sub>ij</sub></i>	172.1	172.1	0.304	8.60	$< 10^{-100}$
<i>TURN<sub>Sij</sub></i>	159.8	159.9	0.353	8.32	$< 10^{-100}$
<i>PADTAL<sub>ij</sub></i>	150.8	150.9	0.390	7.91	$< 10^{-100}$
<i>RAMPS<sub>ij</sub></i>	148.3	148.4	0.400	7.89	$< 10^{-100}$
<i>ROUTE<sub>ij</sub></i>	146.3	146.4	0.408	7.76	$10^{-94}$
<i>PINT<sub>ij</sub></i>	144.6	144.7	0.415	7.61	$10^{-81}$
<i>DISTRO<sub>ij</sub></i>	144.1	144.2	0.417	7.57	$10^{-24}$



**Figure 12. Forward selection criteria versus number of variables for the through trip distribution model.**

The results from the forward selection process suggested that the best model is the three-variable model. However, to allow for the possibility the best model actually includes more variables, all seven variables were retained at this stage of the model development process. These seven variables formed the preliminary model, Model IIa, which is presented in Table 21.

**Table 21. Model IIa**

Variable	Coeff.	Std. Err.	$z$	$p$
$PINTI_{ij}$	1.87	$4.91 \times 10^{-2}$	$3.81 \times 10^1$	$<10^{-4}$
$TURNS_{ij}$	$-3.61 \times 10^{-1}$	$1.17 \times 10^{-2}$	$-3.08 \times 10^1$	$<10^{-4}$
$PADTAL_{ij}$	8.03	$1.88 \times 10^{-1}$	$4.27 \times 10^1$	$<10^{-4}$
$RAMPS_{ij}$	$-2.69 \times 10^{-1}$	$1.39 \times 10^{-2}$	$-1.93 \times 10^1$	$<10^{-4}$
$ROUTE_{ij}$	$8.84 \times 10^{-1}$	$4.66 \times 10^{-2}$	$1.90 \times 10^1$	$<10^{-4}$
$PINT_{ij}$	$9.41 \times 10^{-1}$	$5.38 \times 10^{-2}$	$1.75 \times 10^1$	$<10^{-4}$
$DISTRO_{ij}$	1.45	$1.45 \times 10^{-1}$	9.99	$<10^{-4}$
$\bar{\rho}_0^2 = 0.417$ , Number of responses = 7328				

## FINAL VARIABLE SELECTION

To further investigate each variable in the preliminary model, the model development process used the experiment described in the previous chapter with the same set of study area groups. The results from this experiment are presented in Table 22.

**Table 22. Model IIa Relative  
Parameter Estimate Changes  
(percent)**

Variable	Group Removed			
	1	2	3	4
$PINTI_{ij}$	-1	9	-19	5
$TURNS_{ij}$	-11	4	22	-6
$PADTAL_{ij}$	11	-4	-6	0
$RAMPS_{ij}$	-29	20	2	6
$ROUTE_{ij}$	-10	1	16	-3
$PINT_{ij}$	21	-31	28	-10
$DISTRO_{ij}$	-31	56	-25	-6

The changes in parameter estimates for  $PINTI_{ij}$ ,  $TURNS_{ij}$ ,  $PADTAL_{ij}$ , and  $ROUTE_{ij}$ , are equal to or less than 22 percent, whereas the other three parameters change by at least 29 percent at least once. Thus the variables  $PINTI_{ij}$ ,  $TURNS_{ij}$ ,  $PADTAL_{ij}$ , and  $ROUTE_{ij}$  were retained as the better variables. These variables form Model IIb, which is presented in Table 23.

**Table 23. Model IIb**

Variable	Coeff.	Std. Err.	$z$	$p$
$PINTI_{ij}$	2.60	$3.85 \times 10^{-2}$	$6.76 \times 10^1$	$<10^{-4}$
$TURNS_{ij}$	$-4.79 \times 10^{-1}$	$1.15 \times 10^{-2}$	$-4.17 \times 10^1$	$<10^{-4}$
$PADTAL_{ij}$	7.72	$1.72 \times 10^{-1}$	$4.49 \times 10^1$	$<10^{-4}$
$ROUTE_{ij}$	$8.69 \times 10^{-1}$	$4.71 \times 10^{-2}$	$1.84 \times 10^1$	$<10^{-4}$
$\bar{\rho}_0^2 = 0.397$ , Number of responses = 7328				

## MODEL REFINEMENT

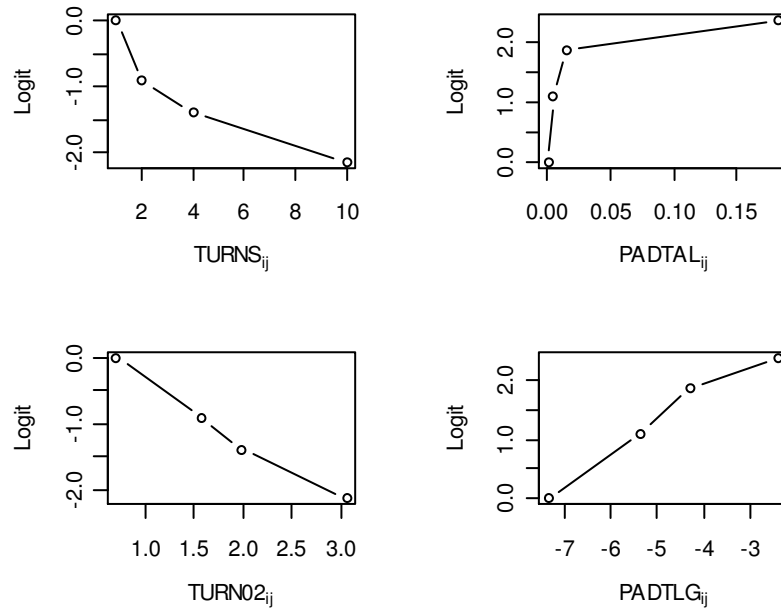
The model development process then tested the assumptions that the continuous variables are linear in the logit. This assumption was tested using the step test as described in the previous chapter. The results of the step test for variables  $TURNS_{ij}$  and  $PADTAL_{ij}$  are presented in Figure 13. Each graph shows evidence that the variable is not linear in the logit. Based on these graphs, two transformed variables were created to replace  $TURNS_{ij}$  and  $PADTAL_{ij}$ , as defined by

$$TURN02_{ij} = \sqrt{TURNS_{ij}}$$

and

$$PADTLG_{ij} = \ln(PADTAL_{ij}).$$

The results of the step test for the new transformed variables are also presented in Figure 13. The plots do not show evidence that the new transformed variables are not linear in the logit, so the new transformed variables were retained to replace the original variables. The model with the transformed variables is Model IIc, and it is the final through trip distribution model. It is presented in Table 24.



**Figure 13. Step test for model IIb continuous variables.**

**Table 24. Model IIc**

Variable	Coeff.	Std. Err.	$z$	$p$
$PINTI_{ij}$	2.24	$3.92 \times 10^{-2}$	$5.72 \times 10^1$	$<10^{-4}$
$TURN02_{ij}$	-1.09	$2.34 \times 10^{-2}$	$-4.65 \times 10^1$	$<10^{-4}$
$PADTLG_{ij}$	$5.72 \times 10^{-1}$	$1.19 \times 10^{-2}$	$4.80 \times 10^1$	$<10^{-4}$
$ROUTE_{ij}$	$8.14 \times 10^{-1}$	$4.77 \times 10^{-2}$	$1.71 \times 10^1$	$<10^{-4}$
$\bar{\rho}_0^2 = 0.407$ , Number of responses = 7328				

The model refinement for the through trip distribution model did not include investigating variable interactions. Interactions with  $ISCV$  are probably the only

potentially useful interactions, but the non-commercial and commercial data was combined to allow the through trip distributions to be based on more observations.

## **MODEL EVALUATION**

The final step in the model development process was to evaluate the performance of the model. This was done using the same procedures as were outlined in the previous chapter for the through trip split model. The cross-classification tables are presented in Table 25 through Table 28. The overall distribution of absolute prediction errors is summarized in Table 29, and the distributions of the prediction errors conditioned on the prediction range are presented in Figure 14 and Figure 15. The absolute prediction error for 62 percent of the observations is less than 1 percent, and only 8 percent of the observations have prediction errors greater than 30 percent (see Table 29), indicating that the model predicts reasonably well.

This evaluation used two sets of four box plots each. The first set of box plots represents only data where the observation was equal to zero. The second set of box plots represents data where the observation was not equal to zero. Separating the data in this way allows a more detailed evaluation of how well the model performs. The first set of box plots is presented in Figure 14. Box plots of the distribution model prediction errors for each quartile of the prediction values (for observations equal to zero). This set of box plots shows that prediction errors when the observation is equal to zero are in general small, but that errors increase as the prediction value increases. Prediction errors are always positive since the observation is always zero.

The second set of box plots is presented in Figure 15. These box plots represent data where the observation is not zero. Again, errors are generally small, and increase with increasing prediction values. Prediction errors tend to be negative.



**Table 25. Predicted versus Observed Through Trip  
Distribution for Group 1**

Observed(%)	Predicted (%)				
	0-1	1-3	3-10	10-30	30-100
0-1	950	260	108	33	1
1-3	17	8	10	3	0
3-10	17	23	32	16	6
10-30	13	15	31	18	9
30-100	4	4	13	11	14

**Table 26. Predicted versus Observed Through Trip  
Distribution for Group 2**

Observed(%)	Predicted (%)				
	0-1	1-3	3-10	10-30	30-100
0-1	739	272	116	44	1
1-3	11	8	6	2	2
3-10	15	9	31	15	2
10-30	6	25	28	21	9
30-100	2	7	11	19	9

**Table 27. Predicted versus Observed Through Trip  
Distribution for Group 3**

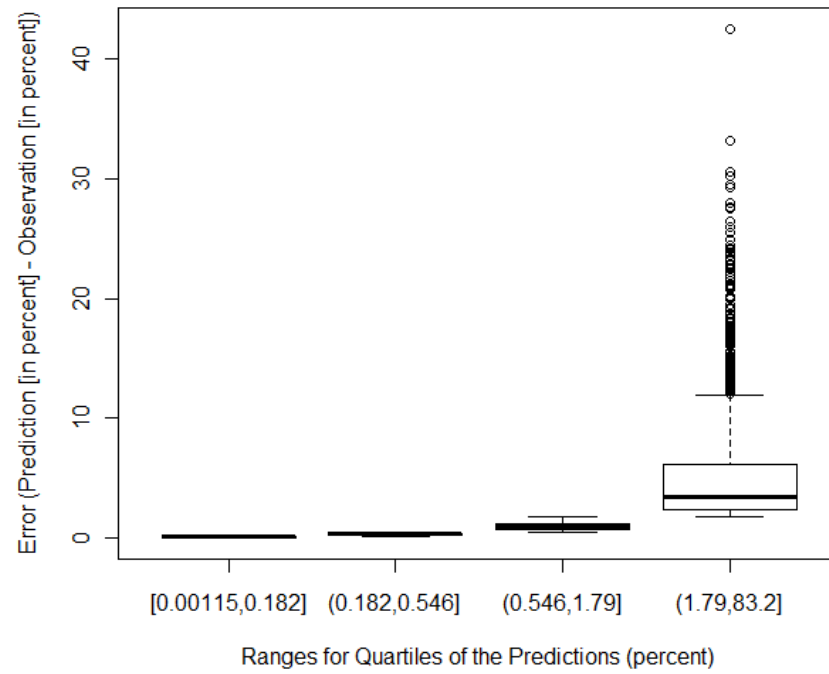
Observed(%)	Predicted (%)				
	0-1	1-3	3-10	10-30	30-100
0-1	3034	664	259	73	1
1-3	42	19	11	7	2
3-10	40	43	42	18	2
10-30	50	32	47	31	4
30-100	7	8	22	25	19

**Table 28. Predicted versus Observed Through Trip  
Distribution for Group 4**

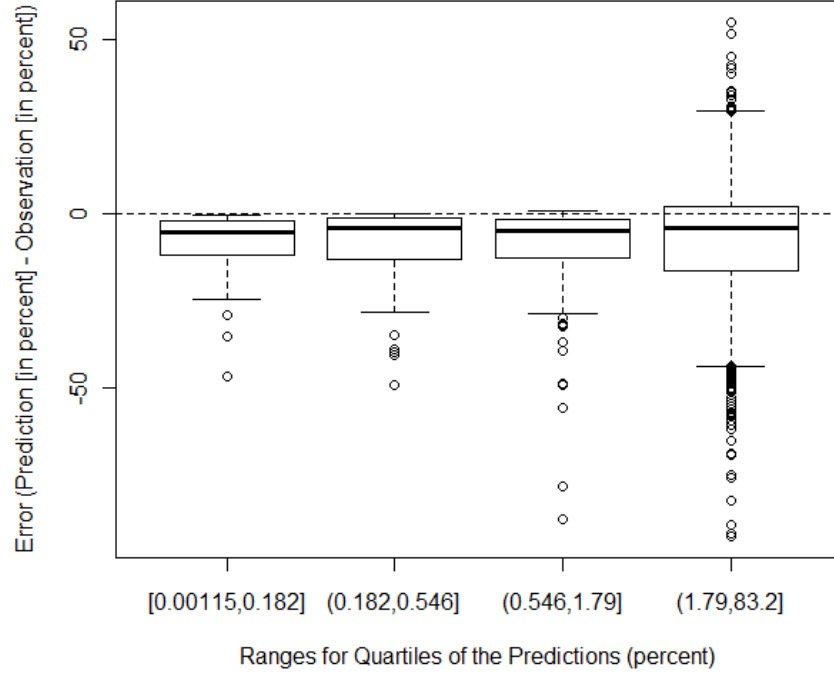
Observed(%)	Predicted (%)				
	0-1	1-3	3-10	10-30	30-100
0-1	440	195	118	24	1
1-3	5	5	1	3	0
3-10	8	14	10	14	1
10-30	4	10	31	15	5
30-100	0	1	7	16	17

**Table 29. Summary of Through  
Trip Distribution Model  
Prediction Errors**

Range of absolute error(%)	Percent of observations
0-1	62
1-3	18
3-10	11
10-30	6
30-100	2



**Figure 14. Box plots of the distribution model prediction errors for each quartile of the prediction values (for observations equal to zero).**



**Figure 15. Box plots of the distribution model prediction errors for each quartile of the prediction values (for observations not equal to zero).**

## RESULTS

This chapter described the model development process for the through trip distribution model, which distributes through trips between external stations. The through trip distribution model requires relatively little data collection, and the model evaluation showed that it predicts reasonably well, both of which indicate that the model can be useful for practical applications. The final form of the model is a multinomial logit model, where the utility of each alternative (entry station)  $i$  for a certain survey station  $j$  is

$$g(\mathbf{x}_{ij}) = 2.24PINT1_{ij} - 1.09TURN02_{ij} + 0.572PADTLG_{ij} + 0.814ROUTE_{ij}$$

where

- $j$  = an index for the external station for which the estimation is made (the survey external station or the external station where through trip exit the study area);
- $i$  = an index for the external station where through trips enter the study area;
- $\mathbf{x}_{ij}$  = the vector of predictor variables for survey external station  $j$  paired with entry external station  $i$ ;
- $g(\mathbf{x}_{ij})$  = the utility of entering the study area at external station  $i$  for a through trip exiting the study area at external station  $j$ ;
- $PINTI_{ij}$  = A binary variable which is 1 if the external station  $j$  has any interaction scores with a route that enters at external station  $i$ ;
- $TURN02_{ij}$  = the square root of number of turns on the least time route between external station  $i$  and external station  $j$ ;
- $PADTLG_{ij}$  = the natural logarithm of the ADT at external station  $i$  as a portion of the total ADT across all possible entry external stations; and
- $ROUTE_{ij}$  = a binary variable which is 1 when the least time route between external station  $i$  and external station  $j$  is valid, and is 0 otherwise.

The inclusion of  $PINTI_{ij}$  in the final model confirms the conclusion drawn from the final form of the through trip split model that the interaction score variables can be important for predicting through trips. An interesting observation is the fact that the forward selection process chooses  $PINTI_{ij}$  as the most important variable for the through trip distribution model, whereas the variable  $PINT_{ij}$  is not included in the final model. The variable  $PINTI_{ij}$  is different from  $PINT_{ij}$  in that it carries no information about the populations of or distances between urban areas. Instead, it simply indicates whether the least time route between two urban areas uses a particular pair of external stations.

The variable  $TURN02_{ij}$  has not been tested in previous research, but it proves to be important for the through trip distribution model. The estimated coefficient is negative, indicating that through trips with few turns are more likely than through trips

with many turns. This is as expected, since turns generally add delay to a trip, and drivers usually try to minimize delay. The square root transformation indicates that the absolute difference in the number of turns between two routes is more important when the numbers are small than when the numbers are large. For example, the difference between zero turns and 1 turn is more important than the difference between 10 turns and 11 turns. In the first case, the route with zero turns will receive significantly more trips than the route with 1 turn. In the second case, the difference in number of trips for the two routes is not as significant.

The variable  $PADTLG_{ij}$  is a transformation of  $PADTAL_{ij}$ , which is included in three of the four stage one models in the literature review, although in this research it is defined somewhat differently. Previous researchers used the ADT as a proportion of the total ADT across all external station. This research excludes the survey station when calculating the total ADT. That (a transformation of)  $PADTAL_{ij}$  is also included in this model confirms that  $PADTAL_{ij}$  is a good predictor of through trips across a variety study areas throughout the nation. The natural logarithm transformation has an effect similar to that of the square root transformation for  $TURN02_{ij}$ ; differences between values of the variable are more important for small values than for large values. For example, the difference between 0.05 and 0.10 is more important than the difference between 0.80 and 0.85.

Like  $TURN02_{ij}$ , the variable  $ROUTE_{ij}$  is new in this research. The estimated coefficient is positive, which indicates that through trips which follow a least-time route are more likely than through trips which follow a non-least-time route. This is expected because drivers normally try to minimize travel time.

## **COMBINED MODEL EVALUATION**

After the development of the through trip split and distribution models was complete, the results from combining the two models was evaluated. These results are the product of the through trip split and through trip distribution predictions. The results were evaluated using the same evaluation methods that were used for the through trip split and distribution models individually. The cross-classification tables are presented in

Table 30 through Table 33. The overall distribution of absolute prediction errors is summarized in Table 34, and the distributions of the prediction errors conditioned on the prediction range are presented in Figure 16. The absolute prediction error for 92 percent of the observations is less than 1 percent (see Table 34), indicating that the combined model predicts well. The box plots confirm that the errors are generally small and show, like the results from the individual models, that the prediction errors increase as the prediction values increase.

**Table 30. Combined Model Predicted versus Observed for Group 1**

Observed(%)	Predicted (%)				
	0-1	1-3	3-10	10-30	30-100
0-1	2849	129	36	1	0
1-3	79	35	19	0	0
3-10	16	17	18	4	0
10-30	4	3	8	9	0
30-100	0	0	2	3	0

**Table 31. Combined Model Predicted versus Observed for Group 2**

Observed(%)	Predicted (%)				
	0-1	1-3	3-10	10-30	30-100
0-1	2467	93	12	0	0
1-3	80	34	12	2	0
3-10	51	22	18	1	0
10-30	8	7	8	2	1
30-100	0	0	2	0	0

**Table 32. Combined Model Predicted versus Observed for Group 3**

Observed(%)	Predicted (%)				
	0-1	1-3	3-10	10-30	30-100
0-1	8326	235	41	2	0
1-3	160	46	22	3	0
3-10	60	29	20	7	0
10-30	8	7	16	16	1
30-100	0	0	0	4	1

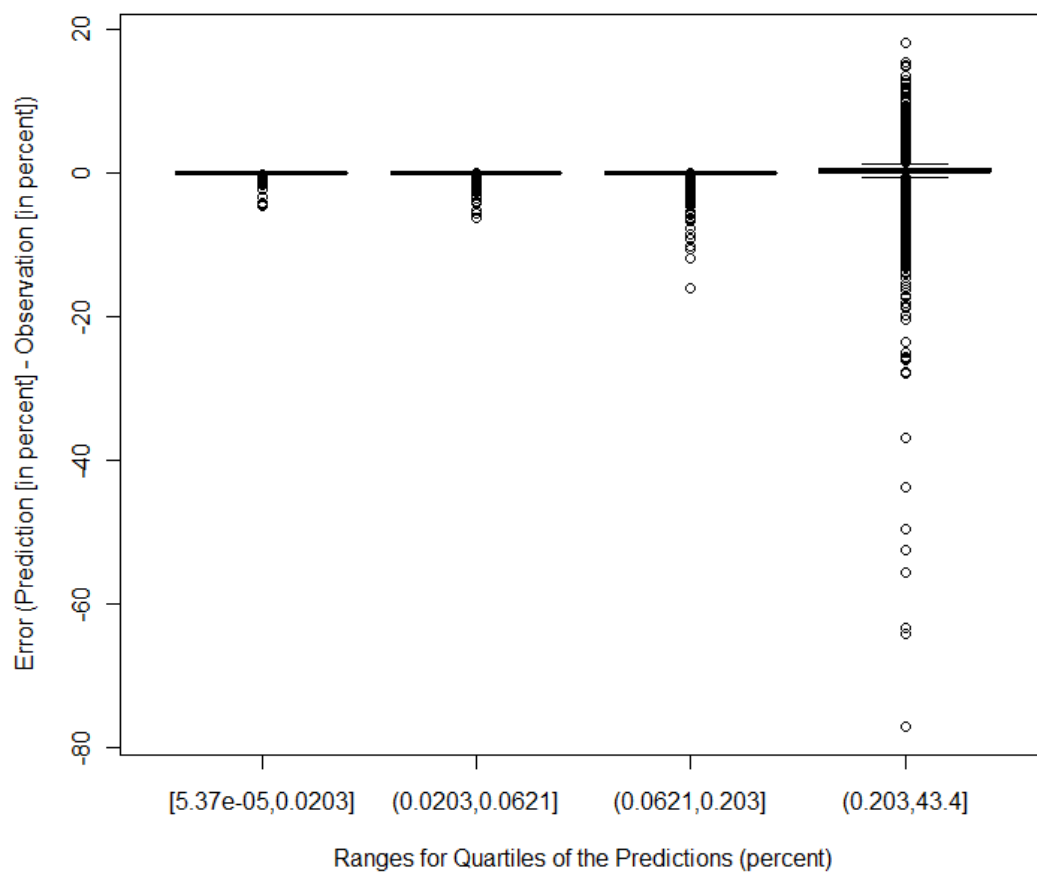
**Table 33. Combined Model Predicted versus Observed for Group 4**

Observed(%)	Predicted (%)				
	0-1	1-3	3-10	10-30	30-100
0-1	1615	88	19	0	0
1-3	43	19	21	1	0
3-10	12	17	13	4	0
10-30	1	2	4	7	0
30-100	0	0	1	3	1

**Table 34. Summary of Combined Model Prediction Errors**

Range of absolute error(%)	Percent of observations
0-1	92
1-3	5
3-10	2
10-30	0
30-100	0





**Figure 16. Box plots of the combined models prediction errors for each quartile of the prediction values.**

## CHAPTER VI

### CONCLUSIONS

This research developed a system of two logit models for estimating through trips. The through trip split model estimates the portion of all trips that are through trips. The specification and parameter estimates for the through trip split model for commercial vehicles are

$$g(\mathbf{x}_j) = -2.94 + 2.38PINTTH + 1.16 - 0.235INTTLI_j$$

and the equation for non-commercial vehicles is

$$g(\mathbf{x}_j) = -2.94 + 2.38PINTTH - 0.235INTTLI_j$$

where

$j$  = an index for the external station for which the estimation is made;

$\mathbf{x}_j$  = the vector of predictor variables for external station  $j$ ;

$g(\mathbf{x}_j)$  = the utility of through trips for vehicles exiting the study area at external station  $j$ ;

$PINTTH_j$  = the proportion through trips at the external station as predicted by the interaction score; and

$INTTLI_j$  = a binary variable which is 1 if the external station has any interaction scores, and is 0 if it has no interaction scores.

The through trip distribution model estimates the portion of through trips that enter the study area at each of the other external stations. The specification and parameter estimates for the through trip distribution model are

$$g(\mathbf{x}_{ij}) = 2.24PINTI_{ij} - 1.09TURN02_{ij} + 0.572PADTLG_{ij} + 0.814ROUTE_{ij}$$

where

$j$  = an index for the external station for which the estimation is made (the survey external station or the external station where through trip exit the study area);

- $i$  = an index for the external station where through trips enter the study area;
- $\mathbf{x}_{ij}$  = the vector of predictor variables for survey external station  $j$  paired with entry external station  $i$ ;
- $g(\mathbf{x}_{ij})$  = the utility of entering the study area at external station  $i$  for a through trip exiting the study area at external station  $j$ ;
- $PINTI_{ij}$  = A binary variable which is 1 if the external station  $j$  has any interaction scores with a route that enters at external station  $i$ ;
- $TURN02_{ij}$  = the square root of number of turns on the least time route between external station  $i$  and external station  $j$ ;
- $PADTLG_{ij}$  = the natural logarithm of the ADT at external station  $i$  as a portion of the total ADT across all possible entry external stations; and
- $ROUTE_{ij}$  = a binary variable which is 1 when the least time route between external station  $i$  and external station  $j$  is valid, and is 0 otherwise.

The models were developed using data from study areas with populations ranging from 100,000 to 6 million people, so it can be applied to larger study areas than can previously developed models. The model system produces reasonably accurate estimates. For the through trip split model, 49 percent of the absolute errors are less than 5 percent, and only 10 percent of the absolute errors are more than 20 percent. For the through trip distribution model, 62 percent of the absolute errors are less than 1 percent, and only 8 percent of the absolute errors are more than 10 percent. The model system requires little data or time, so it is practical for application in Texas study areas.

In addition to developing the two models, this research also led to two conclusions that help in understanding travel patterns in general. The first conclusion of this research is that any estimation of through trips must take into account the location of the study area boundary, since the boundary alone determines which trips are through trips and which are not. In some cases, the location of the study area boundary can provide information that can be used for through trip estimation. For example, if the

study area boundary is located in such a way that all external stations are on highways that provide direct access to the urban area within the study area, then a reasonable assumption is that the characteristics of the urban area can be used to predict through travel. On the other hand, if the study area boundary is located arbitrarily or randomly, then the location of study area boundary provides less information that can be used to estimate through trips.

This conclusion implies that MPOs can improve their through trip estimations (or reduce data requirements for a certain quality of through trip estimations) by carefully choosing study area boundaries. MPOs should attempt to (as much as possible) choose the study area boundary such that:

- all external stations are on highways that provide direct access to the main urban area; and
- the least time route between any two external stations crosses the boundary only at the two external stations and passes through the study area; and
- the least time route between any external station and any location inside the study area crosses the study area boundary only at the external station.

Choosing the study area boundary in this way will generally result in better through trip estimates with the same data, because the location of the boundary provides additional information that can be used in the estimation.

The second conclusion of this research is that one of the most important predictors of through travel is the regional context of the study area, meaning the geographic location of highways and urban areas surrounding the study area, as evidenced by the inclusion of interaction score variables in the two models. In fact, this type of information is so valuable that reasonably accurate through trip estimates could be made even for a study area of random size, shape, and location, given that the data has adequate detail and scope. Extending the data collection requirements to areas outside of the study area will increase the cost of the estimation, but this research shows

that even an extremely simple model of travel in the region surrounding the study area contributes significantly to estimating through trips. MPOs should consider developing simple regional models to improve through trip estimations for the study area.

The model development process also highlighted an area of potential future research. The number of turns proved to be a very important predictor of the distribution of through trips. This is interesting since the number of turns provides no information about the values normally used to distribute trips, which are trip duration and distance. Future research should investigate the possibility of using the number of turns as a way to describe and estimate travel behavior in general.

## REFERENCES

- Anderson, M. D. (2005). "Spatial economic model for forecasting the percentage splits of external trips on highways approaching small communities." *Transp. Res. Rec.*, 1931, 68-73.
- Anderson, M. D. (1999). "Evaluation of models to forecast external-external trip percentages." *J. Urban Plan. Dev.*, 125(3), 110-120.
- Anderson, M. D., Abdullah, Y. M., Gholston, S. E., and Jones, S. L. (2006). "Development of a methodology to predict through-trip rates for small communities." *J. Urban Plan. Dev.*, 132(2), 112-114.
- Cambridge Systematics, Inc. (1996). *Quick Response Freight Manual*. Federal Highway Administration, Washington, D.C.
- Chatterjee, A., and Raja, M. (1989). "Synthetic models for through trips in small urban areas." *J. Transp. Eng.*, 115(5), 537-555.
- Denver Regional Council of Governments, and Parsons Transportation Group, Inc. (2000). *Denver Regional Travel Behavior Inventory: Household Survey Report*, <<http://www.drcog.org/index.cfm?page=RegionalTravelBehaviorInventory>> (May 20, 2010).
- Greene, W. H. (n.d.). *Limdep Version 9.0 Econometric Modeling Guide*. Econometric Software, Inc., Plainview, NY.
- Han, Y. (2007). "Synthesized through trip model for small and medium urban areas." Ph. D. dissertation, North Carolina State Univ., Raleigh, N. Ca.

- Han, Y., and Stone, J. R. (2008). "Synthesized through-trip models for small and medium urban areas." *Transp. Res. Rec.*, 2077, 148-155.
- Holguín-Veras, J., and Patil, G. (2005). "Observed trip chain behavior of commercial vehicles." *Transp. Res. Rec.*, 1906, 74-80.
- Horowitz, A. J., and Patel, M. H. (1999). "Through-trip tables for small urban areas: a method for quick-response travel forecasting." *Transp. Res. Rec.*, 1685, 57-64.
- Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*, Wiley, New York.
- Huff, D. (1963). "A probabilistic analysis of shopping-center trade areas." *Land Economics*, 39(1), 81-90.
- Koppelman, F. S., and Bhat, C. (2006). *A self instructing course in mode choice modeling: multinomial and nested logit models*, < [http://www.civil.northwestern.edu/people/koppelman/PDFs/LM\\_Draft\\_060131Final-060630.pdf](http://www.civil.northwestern.edu/people/koppelman/PDFs/LM_Draft_060131Final-060630.pdf) > (May 20, 2010)
- Martchouk, M., and Fricker, J. D. (2009). "Through-trip matrices using discrete choice models: planning tool for smaller cities." 88<sup>th</sup> *Annual Meeting Compendium of Papers*, Transp. Res. Board, Washington, D.C.
- Martin, W. A., and McGuckin, N. A. (1998). *Travel Estimation Techniques for Urban Planning*. NCHRP Report 365, Transp. Res. Board, Washington, D.C, 170.
- Modlin Jr., D. G. (1974). "Synthetic through trip patterns." *J. Transp. Eng. Division*, 100(2), 363-378.

- Modlin Jr., D. G. (1982). "Synthesized through-trip table for small urban areas." *Trans. Res. Rec.*, 842, 16-21.
- Pigman, J. G., and Deen, R. C. (1979). "Simulation of travel patterns for small urban areas." *Trans. Res. Rec.*, 730, 23-29.
- Reeder, P. R. (1993). "Comparison of synthetic thru trips to recent external O&D data." *Fourth National Conference on Transportation Planning Methods Applications*, Transp. Res. Board, Washington D.C.
- Spiess, H. (1987). "A maximum likelihood model for estimating origin-destination matrices." *Transp. Res. Part B*, 21(5), 395-412.
- Zipf, G. (1946). "The (P1 P2)/D hypothesis: on the intercity movement of persons." *American Sociological Review*, 11(6), 677-686.



### **VITA**

Eric Talbot received his Bachelor of Science degree in civil engineering from Brigham Young University in 2007, and graduated with his Master of Science degree in civil engineering from Texas A&M University in 2010. Mr. Talbot may be reached at: Resource Systems Group, Inc., 55 Railroad Row, White River Junction, VT 05001.